이학박사학위논문

유전체 재분석을 통한 배추 근연속 식물 7 종들에 대한 다양성 및 진화에 관한 연구

Genomic diversity and evolution of seven *Brassica* species revealed by whole genome resequencing

2016 년 8 월

서울대학교대학원

생물정보학협동과정

설 영 주

Genomic diversity and evolution of seven *Brassica* species revealed by whole genome resequencing

by Young-Joo Seol

Advisor: Professor Jongsik Chun, Ph.D.

A Thesis Submitted for the Partial Fulfillment of the Degree of Doctor of Philosophy

AUGUST 2016

Interdisciplinary Graduate Program in Bioinformatics
Seoul National University

ABSTRACT

The *Brassica* genus contains the most diverse collection of agronomically important plant species and is a relative of the model plant *Arabidopsis thaliana*, from which it diverged ~20 MYA. The six most agro-economically important *Brassica* species include the three diploid species, *Brassica rapa* (AA, 2n = 20), *Brassica oleracea* (CC, 2n = 18), and *Brassica nigra* (BB, 2n = 16); and the three allotetraploid species, *Brassica juncea* (AABB, 2n = 34), *Brassica napus* (AACC, 2n = 38), and *Brasscia carinata* (BBCC, 2n = 36), which were formed through the hybridization of their diploid genome counterparts. To understand genetic relationship and evolution of the U's triangle in *Brassica* species, whole genome resequencing of 28 *Brassica* species belonging to A, B, C, R, AB, AC, and BC genome was conducted using Illumina MiSeq next-generation sequencing platform. Approximately 6~8 million sequence reads were obtained from each genotype and ~80% of them were high quality sequences. Overall, ~87% of the total pre-processed sequence reads from each genotype were mapped to publicly available multiple reference genomes including *A. thaliana, B. rapa, B. oleracea,* and *B. napus*. The average mapping depth was over three-fold for each genotype, and 59 million high-confidence genome-wide single-nucleotide polymorphisms (SNPs) and 247,407 indels were detected across the reference genomes. Using *Arabidopsis* as a reference, variants derived from exon were much higher than that of from intergenic or intron, which suggested that intergenic and intron sequences went divergent faster after *Arabidopsis* and *Brassica* species split. In comparison of all four kinship analysis based on SNPs with multiple references (*A. thaliana, B. rapa, B. oleracea* and *B. napus*), each of the four accessions representing a diploid genome type was grouped in a cluster regardless of reference genome. The position of allotetraploid

genomes in phylogenic tree was incongruent due to the complex history of *Brassica* lineage when multiple *Brassica* references were used, however the genetic relationship including allotetraploid wasmore clearly explained with higher quality SNPs using *A. thaliana* as a reference.

In order to reveal the diversity, origin and evolution of *Brassica* species, additionally a comprehensive phylogenetic analysis of 28 *Brassica* species was carried out based on complete chloroplast (CpDNA) and 45S ribosomal sequences (nrDNA). Concurrent phylogenomic analysis elucidates the genetic diversity, relationship, maternal source for allotetraploids and evolution of the *Brassica* species. In addition, complete map of the structural variants such as SNPs, indels, and copy number variations for CpDNA and nrDNA were constructed. An independent estimation of divergence time, based on CpDNA and nrDNA together with previous reports reveals the allotetraploids were diverged about 0.01 MYA. Structural variants such as SNP and indel have provided potential barcoding markers for identification of each *Brassica* species including *Raphanus sativus*. Certainly, the use of CpDNA and nrDNA provides a comprehensive overview of the genome diversity and evolutionary context of the major *Brassica* species.

Keywords: Whole genome resequencing, NGS, SNPs, diversity and evolutionary relationship, CpDNA, nrDNA, *Brassica*

Student number: 2009-30107

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

BAC	Bacterial artificial chromosome
BLAST	Basic local alignment search tool
bp	Base pair
DNA	Deoxyribonucleic acid
EST	Expressed sequenced tags
GATK	Genome analysis toolkit
indel	Insertion/deletion
MYA	Million years ago
MY	Million years
NCBI	National center for biotechnology information
NGS	Next-generation sequencing
PCR	Polymerase chain reaction
qRT-PCR	quantitative real-time PCR
QTL	Quantitative trait loci
SNP	Single nucleotide polymorphism
SSR	Simple sequence repeat
SV	Structural variation

Chapter I. General Introduction

The U's triangle model for explaining the relationship among *Brassica* crops

The genus *Brassica* (Brassicaceae) is agriculturally important and one of the largest family (340~400 genera and approximately 3709 species) in the plant kingdom. Brassicaceae includes many economically valuable crop species, with wide range of morphological diversity including chinese cabbage, mustard, cauliflower, broccoli, turnip, radish, and oilseed rape (Bailey et al. 2006; Al-Shehbaz et al. 2006; Paterson et al. 2001). The Brassicaceae family encompasses extensive species diversity with a wide range of intra- and inter-specific morphological and phytochemical profiles, which has contributed to their global importance in crop production. The genomic relationship of six interrelated *Brassica*s was described in the U's triangle (Figure 1.1) (Johnston et al. 2005) includingthree diploids - *Brassica rapa* (2*n* :20, AA, 529 Mb), *B. nigra* (2*n* :16, BB, 632 Mb) and *B. oleracea* (2*n* :18, CC, 696 Mb), and three amphidiploid derivatives - *B. juncea* (2*n* :36, AABB, 1068 Mb), *B. napus* (2*n* :38, AACC, 1132 Mb) and *B. carinata* (2*n* :34, BBCC, 1284 Mb). Comparative study of *Brassica* species with the diploid model plant, *A. thaliana,* confirmed that approximately 16-fold variation in genome size was established in *Brassica* family. In addition, comparative analysis revealed that *B. rapa* and its close relative *B. oleracea* evolved as triploid derivatives from the common ancestor,*A. thaliana,* from 4-17 MYA (Mun et al. 2009; Koenig and Weigel 2015).

Importance of studying *Brassica* genus

Brassica species form an important source of vegetable oil, fresh and preserved vegetables, and condiments. About 76 million tons of *Brassica* vegetables were produced in 2010 with a value of 14.85 billion dollars (http://faostat.fao.org/). Due to its wide geographical distribution and the

different polyploidy nature, the family Brassicaceae provides an excellent material to study the evolution of the plant polyploidy(Koch and Kiefer 2006).

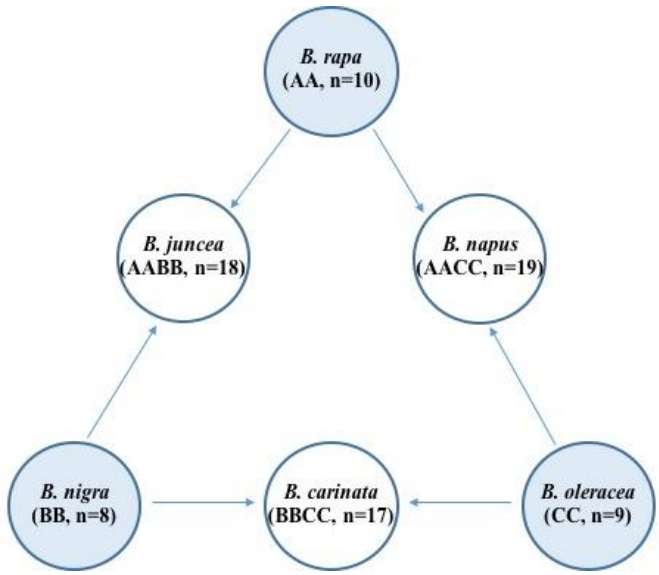

Figure 1. Genomic relationships among six cultivated *Brassica* species represented by U's triangle. (U, 1935; Johnston et al. 2005).

In particular, Chinese cabbage (*B. rapa ssp. pekinensis*) is one of the most widely consumed vegetable crops in northeast Asia. Because of the high commercial value of *Brassica* throughout the world, their agricultural characteristics have been constantly targets for crop improvement. *Brassica* species are characterized by a remarkable morphological diversity with regard to inflorescences, leaves, stems, roots and apical buds (Paterson et al. 2001). Morphological diversity exists even within a species including the enlarged inflorescence of cauliflower (*B. oleracea ssp. botrytis*) and broccoli (*B. oleracea ssp. italica*); the enlarged stem of kohlrabi(*B. oleracea ssp. gongylodes*) and marrowstem kale (*B. oleracea ssp. medullosa*); the enlarged root of turnip (*B. rapa ssp. rapifera*); the enlarged and twisted leaves of Pak-choi (*B. rapa ssp. chinesis*) and Chinese cabbage (*B. rapa ssp.*

pekinesis); and the enlarged single apical bud of cabbage (*B. oleracea ssp. capitata*) or the many axillary buds of Brussels sprout (*B. oleracea ssp. gemmifera*). Such phenomenon has been linked to both recent and ancient polyploidy genomic changes. In *Brassica* species, such polyploidization events have induced genome triplication and rearrangements accompanying genetic variation, including insertions, deletions and substitutions. These results have effected novel phenotypic variations for important traits (Lukens et al. 2004; Yang et al. 2006; Town et al. 2006; Park et al. 2005).

The rich diversity of *Brassica* plants

B. rapa(AA), one of the diploid *Brassica*, is a major vegetable or oil crop in Asia and Europe. It has recently become a widely used model to study the polyploid genome structure and evolution because of its smallest genome size (529 Mb) among the *Brassica* genus. This species has evolved from a hexaploid ancestor aimilar with all members of the tribe Brassicaceae (Johnston et al. 2005; Yang et al. 2006; Mun et al. 2009). The diploid *B. nigra* (BB) has not been studied extensively at the genomic level compared to other *Brassica* species despite the research on agronomically important genes in terms of disease resistance, drought tolerance, and seed oil quality (Chevre et al. 1996; Pakpour and Klironomos 2015; Struss et al. 1996). The other diploid *B. oleracea* (CC) comprises many important vegetable crops including cauliflower, broccoli, cabbages, bBrussels sprouts, kohlrabi, and kales. These species demonstrate extreme morphological diversity in terms of leaves, flowers, and stems. Most of the *B. oleracea* are rich in proteins and carotenoids (Kopsell and Kopsell 2006), and diverse glucosinolates (GSLs) that function as unique phytochemicals for plant defense against fungal and bacterial pathogens (Halkier and Gershenzon 2006). Furthermore, consumption of *B. oleracea* has shown to have potential anticancer effects (Beecher 1994; Hafidh et al. 2013).

The rapeseed (*B. napus*, AACC, 2n = 4X = 38) is an important vegetable and oilseed crop grown in different parts of the world, such as Asia, Europe, and North America. It was adapted in wide geographical regions in various forms and this has apparently enhanced genetic diversity in this species. It is an allotetraploid species in the Brassicaceae family and was resulted from the natural hybridization between *B. rapa* L. (2n = 20, AA) and *B. oleracea* L. (2n = 18, CC) about ~7500 years ago, followed by chromosome doubling, a process known as allopolyploidy and artificial domestication (approximately 400–500 years). These three species and *Arabidopsis* are thought to share a common ancestor (Qian et al. 2006; Chalhoub et al. 2014; Venglat et al. 2013; Hua et al. 2012). *B. juncea* (AABB) is one of the six major cultivated *Brassica* and an important edible oil crop in India. It accounts for nearly 30% of the total oilseed production and 31.4% of an edible oil pool of the country (Singh et al. 2014). Like *B. napus*, the genetic and genomic studies in *B. juncea* have been performed less extensively. However, in recent years, scientists have given more attention to *B. juncea* because of its resistance to salinity and seed shattering (Shekhawat et al. 2012).

Recent work on the amphidiploids, *B. carinata* (BBCC) outperformed on its adaptability and productivity in semi-arid and temperate regions compared to oilseed rape. Because of resistance to various diseases and biotic stress, *B. carinata* is suitable to cultivate in temperate environments and also a potential crop for biofuel production (Cardone et al. 2003; Gaur and Meena 2016). Although genetic diversity analysis of this species has been extensively performed, limited work has been done at genomic levels. Radish (*Raphanus sativus* L.) is an annual root vegetable crop cultivated worldwide and has a substantial role in seed industry, especially in EastAsia. It also serves as an excellent model system to study polyploidy-related genome evolution because of its paleohexaploid ancestry and its close evolutionary relationship with *Arabidopsis* and other *Brassica*

species (Lukens et al. 2004). *R. sativus* was included in this study for genome comparison with species of *Brassica* and *A. thaliana,* one of the close relatives of the model organism.

Genomic resources of *Brassica* species

Current developments in genome sequencing shows significant progress on increased throughput accompanied by plunging costs. High-throughput sequencing technologies are now routinely applied to a wide range of whole-genome sequencing projects in non-model organisms and comparative genomics to address important biological questions that were not possible before(Michael and Jackson 2013). To date, the genomes of ten *Brassica* species have been partially or completely sequenced, including the model plant *Arabidopsis* (Table 1). The annotated *Arabidopsis* genome sequence provides a valuable reference to develop DNA markers and linkage maps to identify candidate genes in cultivated *Brassica* species. Most of the ancestral progenitor sequences used for genome evolution studies and identification of conserved ancestral genomic segments. The sequencing project of 1,001 accessions of *Arabidopsis* will enable genome-wide association in this species (http://1001genomes.org)by linking phenotypic diversity with genetic diversity. The large genomics resources generated from *Arabidopsis* 1001 project will be a foundation for various scientific investigations and breeding applications.

The genomes of *B. rapa* and two sister species, *B. oleracea,* and *B. napus,* have been sequenced recently (Chalhoub et al. 2014) and eight other Brassicaceae species have also been sequenced (Hu et al. 2011; Haudry et al. 2013; Slotte et al. 2013; Kagale et al. 2014)(Table 1). These Brassicaceae genome datasets are a valuable resource for genome and gene studies among the closely related Brassicaceae species. Whole genome sequencing data of *Brassica* species is valuable to elucidate the important genes, understand the

genome evolution and improvecrop quality. Sequencing of the *Brassica* genomes was initiated in 2002 by the Multinational *Brassica* Genome Project (MBGP). The Chinese cabbage, *B. rapa* (cv. Chiifu- 401), was the first genome selected for sequencing because of its small genome size (529 Mb) and low frequency of repetitive sequences. The draft genome sequence of *B. rapa* (A genome) consisted of 10 pseudo-chromosomes with a total of 41,174 protein-encoding genes (Wang et al. 2011b). However, *B. rapa* genome is still rapidly changing. There are two least factors that drive change in the *B. rapa* genome. First, as a species with relatively recent whole genome triplication, the *B. rapa* genome is still experiencing gene fractionation (Mun et al. 2009). Second, the transposons in the *B. rapa* is very active (Mun et al. 2009). Both of these factors create large number of variations within the species. The genome sequence of another important vegetable crop, *B. oleracea* (C genome), has recently been completely determined using a whole genome shotgun (WGS) sequencing strategy. A 630 Mb assembled draft genome sequence was obtained, with a scaffold N50 size of 1.457 Mb and contig size of 26.828 Kb, and assigned to nine pseudo-chromosomes containing 45,758 predicted genes (Liu et al. 2014; Yu et al. 2013).

Recently, the genome of *B. napus* (1130 Mbp), a polyploid genome has been sequenced. *B. napus* originated from a recent combination of two distinct genomes approximately 7500 years ago and gave rise to the crops of rape oilseed (canola), kale and rutabaga. The genome assembly covers ~79% of estimated genome size and includes 95.6% of *Brassica* expressed sequence tags (ESTs). The assembled C_n subgenome (525.8 Mb) is larger than the A_n subgenome (314.2 Mb), consistent with the relative sizes of the C genome of *B. oleracea* (540 Mb, 85% of the ~630Mb genome) and the A genome of *B. rapa* (312 Mb, 59% of the ~530-Mb genome). Furthermore, a total of 45,758 protein-coding genes were predicted, with a mean transcript length of 1,761 bp, a mean coding length of 1,037 bp, and a mean of 4.55

exons per gene(Chalhoub et al. 2014). Currently, the whole genome sequencing of other *Brassica* and *Raphanus* species is in progress and will be completed in the near future(Sharma et al. 2014).

Table 1. Summary of genome sequences completed in Brassicaceae species

Species	Genome size (Mb)	No. of predicted genes	% of genes orthologous to *A. thaliana*	References
Aethionema arabicum	240	23,167	72.4	(Haudry et al. 2013)
Arabidopsis lyrata	230–245	27,379	92	(Hu et al. 2011)
Arabidopsis thaliana	125	28,710	100	*Arabidopsis* Genome Initiative 2000
Brassica napus	849.7	91,167	-	(Chalhoub et al. 2014)
Brassica oleracea	696	45,758	-	(Liu et al. 2014)
Brassica rapa	529	41,174	78.2	(Wang et al. 2011a)
Capsella rubella	210–216	26,521	88	(Slotte et al. 2013)
Eutrema salsugineum	314	26,521	82.7	(Yang et al. 2013)
Leavenworthia alabamica	316	30,343	67.7	(Haudry et al. 2013)
Schrenkiella parvula	140	28,901	80.2	(Dassanayake et al. 2011)
Sisymbrium irio	262	28,917	82.9	(Haudry et al. 2013)

The improvement of sequencing technology has provided vast genomic information and sequence data for majority of the crop species. In the past decade, various *Brassica* databases were integrated on a common platform to facilitate efficient utilization by diverse researchers. An open access integrated database provides annotated genome information, genetic and physical maps, molecular markers, reference maps and gene expression data. The UK *Brassica* community put initiative in this direction in 1996 by compiling *Brassica* sequences and genetic maps to create the BrassicaDB database. A major advancement in knowledge sharing realized the initiation of the Multinational *Brassica* Genome Project (MBGP) in 2002. A number

of open access databases are available in Brassicaceae with various information on linkage maps, QTL maps, details of mapping populations, BAC libraries, marker data, EST repositories and genome sequences. The annotated *B. rapa* genome sequence is available on BRAD *Brassica* database (Cheng et al. 2011) and Brass ensemble (Rothemsted Research, UK) web resources. The *B. oleracea* genome sequence has been available for comparative analysis on the Bolbase data source (http://www.ocri-genomics.org/bolbase/index.html), although the complete genome for download is yet to be released. RadishBase, a database of genetics and genomics of radish, was recently developed by Cornell University(USA) and consists of SSR, EST, and SNP marker information, linkage maps, and organelle genome sequences. Currently, many genetic and genomic resources in Brassicaceae are available and are summarized in Table 2. The integrated knowledge including genomics, transcriptomics, metabolomics, and even phenomics which are available in the public domain, will provide a platform to exchange information and a basis for *Brassica* breeding improvement.

Table 2. List of genetic and genomic web resources for Brassicaceae species

Resource	URL	Remarks
ACPFG	http://www.Brassicagenome.net/	*B. rapa* genome browser, EST-SNP data base, BrassicaDB, CMap to compare genome and genetic map
Bolbase	http://www.ocri-genomics.org/bolbase/	Genomic data of *B. oleracea*, analysis of genome structureas well as syntenic regions, browse, search and downloadgenome of *B. rapa* and *A. thaliana*
BRAD	http://Brassicadb.org/brad/	Compilation of sequence datasets including the completesequence of *B. rapa*.
BrassEnsembl	http://www.Brassica.info/ BrassEnsembl/index.html	*B. rapa* genome sequence, consensus integrated genetic maps of the *Brassica* A and C genomes
BrassicaGenome Gateway	http://Brassica.nbi.ac.uk	*Brassica* genome sequencing database, *Brassica* 95K unigenes set, the *Brassica* IGF Project, *Brassica*DB
Brassica.info	http://www.Brassica.info/	Web-based open source to exchange information relating to *Brassica* genomics and genetics and registries of reference datasets
BrassicaDB	http://Brassica.nbi.ac.uk/BrassicaDB/	Comprehensive sequence data set, genetic maps and markersin *Brassica* species, BLAST server, physical maps
CropStoreDB	http://www.cropstoredb.org	A collection of datasets related to plant and crop genetics, *Brassica* data implemented
Genoscope	www.genoscope.cns.fr/Brassicanapus/	*Brassica napus* Genome Browser
PlantGDB	www.plantgdb.org/BrGDB/	Assembled and annotated*B. rapa* genome sequence
Radish database	http://radish.plantbiology.msu.edu/index.php/Main_Page	EST sequences, linkage maps, SNP and SSR markers, radishgenome sequence updates
RadishBase	http://bioinfo.bti.cornell.edu/cgi-bin/radish/index.cgi	Assembled and annotated ESTs, predicted metabolic pathways, EST-SSR, SNP markers, and genetic maps

Chloroplast genomes of *Brassica* genus

The chloroplast (cp) genome contains abundant information shaped by speciation and it is a rich resource to trace evolutionary processes in population and divergence. Therefore, the cp genome sequence is very important in several fields of plant biology, including phylogenetics, molecular biology, evolutionary biology, and cp genetic engineering. Complete sequences of a tobacco and a liverwort cp genomes were first reported in 1986 (Ohyama et al. 1986; Shinozaki et al. 1986). Since then, cp genomes from a number of land plants and algae have been determined. Because of the characteristics of conserved genome size, gene arrangement, and coding sequences among cp genomes, a PCR based approach has been used for their amplification, sequencing, and assembly (Cronn et al. 2008). However, the development of next-generation sequencing technologies has shed new light on assembly of complete cp genomes. Pyrosequencing of angiosperm plastid genomes were the first attempt to use second generation sequencing technology (454 GS) for the cp genome (Moore et al. 2006).

DNA barcoding for characterizing species

An accurate classification of a large number of species remains a noticeable problem for not only ordinary scientist but also taxonomists. The emergence of DNA barcode has made a positive impact on biodiversity classification and identification (Gregory, 2005). DNA barcoding is a way distinguishing a species using a short DNA sequence derived from a standard position in a genome (http://barcoding.si.edu/DNABarCoding.htm). The possibility of using the chloroplast genome as a ‘super-barcode’ was assessed and the concept of a ‘specific barcode’ derived from comparisons among plastid genome sequences in a targeted group of taxa promised as an

effective alternative that might be widely used for plant identification studies. Specific barcodes can bring new dimensions in the quest for rapid and reliable species discrimination, mainly within closely related plants. At present, DNA barcoding technology depends heavily on chloroplast sequences because of their slightly slow evolutionary process compared to nuclear loci (Dong et al., 2012). The full cp-genome has a relatively conserved sequence content, which ranges from 110 to 160Kb. Currently genic regions in cp are primary targets to design DNA barcode however, cp comparisons in full-lebgth can provide more variations to distinguish closely related plants and even accessions in a species. It can significantly increase resolution at lower taxonomic levels in plant phylogenetic, phylogeographic and population genetic analyses. This super-barcode in turn helps in the recovery of lineages as monophyletic hence it is a species-level DNA barcode (Parks et al., 2009).

The use of cp-genome as a marker prevents possible issues with gene deletion and low PCR efficiency (Huang et al., 2005). The analysis of super-barcode again eliminates the problems associated with sequence retrieval that usually appear in traditional barcoding studies. Small size and higher interspecific and lower intraspecific divergence compared to the nuclear genome makes cp genome most suitable as a genome based barcode. Super-barcoding is more efficient in detecting gene loss and defining gene order than the classical barcoding thus, it promise to provide a faster and simpler means in species identification (Hebert et al., 2004; Luo et al., 2008, 2009).

A complete cp-genome assembly using whole–genome shotgun sequence data is an accurate and less resource intensive than that of using purified chloroplast DNA (McPherson et al., 2013). Therefore, both extraction methods and sequencing capacity are no longer hindrances to obtaining complete cp-genome in many plants and this advanced *de novo* cp assembly will promote to produce many individual super-barcodes

(Doorduin et al., 2011). The complete chloroplast sequence as a super-barcode will finally provide resources to study genetic relationships and diversity in inter- and intra-specific plant groups (Bayly et al., 2013; Yang et al., 2013)

Research objective

As described above, comparative sequence analysis across the species has been utilized as a method to understand genome structure, evolution, and the detection of conserved genomic segments. This is an important field to study genome evolution, sequence collinearity, and transfer of information from extensively studied model organisms to species of commercial interest. Hence, the objective of this research presented here is to elucidate the diversity and evolution of seven *Brassica* species using re-sequencing data of 28 *Brassica* accessions (Table 3). The genome wide variation in *Brassica* species was investigated by mapping 28 re-sequenced data to reference genomes of *B. rapa*, *B. oleracea*, and *B. napus*, and additionally, the diversity and evolution of the *Brassica* species were studied among those species in U's triangle. In addition, the complete chloroplast genome sequences of 28 *Brassica* species were generated using *de novo* assembly, and phylogenetic relationship and chloroplast structure were also analyzed. Lastly, consensus/specific markers for *Brassica* genus were designed from the chloroplast genome resources. Holistically, this study did not only provide insights into *Brassica* genome evolution but also underpin research into the many important crops in this genus and it becomes helpful resources to the *Brassica* breeders and geneticists. It is hoped that this modest compendium marks the beginning of a vibrant future for *Brassica* comparative genome biology, gene discovery, molecular marker developmet, and genetic dissection of important traits.

Chapter II. SNP based genome wide comparison analysis

INTRODUCTION

The genus *Brassica* is the most important vegetable crop, and is consumed daily as food worldwide due to its nutritional values for human. Besides its economic importance, *Brassica* species are considered a model plant for studying genome evolution of the plant polyploidy (Koch and Kiefer 2006). Recently, numerous whole genome sequences of *Brassica* species were reported (Table 1) and the availability of these reference genomes enhances our understanding of genome architecture, and evolution of *Brassica* species, as well as facilitates identification of genes associated with important traits for crop improvements.

The advent of next-generation sequencing (NGS) and bioinformatics tools have revolutionized the field of molecular biology and specifically genomics. It is now possible to generate large amounts of sequence data for answering biological questions very rapidly and at substantially with lower costs. NGS can be employed to a wide range of applications including *de novo* genome, transcriptome, epigenome, non-coding RNA discovery, molecular marker, gene discovery, comparative and evolutionary genomics, and association studies (Sharma et al. 2014). The diverse applications of NGS attracts a broader scientific community in order to answer the complex and fundamental biological questions utilizing and interpreting high-throughput big data, which is based on well-designed but complex experiments.

A complete and well-annotated reference genome sequence provides the ultimate answers for genomic and genetic questions. Recently, whole-genome resequencing of genetic stocks or germplasm becomes popular for the species that a reference genome is available. Because a genome sequence carries the most complete information of genetic variations [e.g. structural rearrangements, copy number variation, insertion–deletion, single nucleotide polymorphisms (SNPs) and sequence repeats],

resequencing of many individuals in a population will be the method for genetic studies to understand the genetic diversity and discover useful alleles. Resequencing can expedite the identification of genetic variations within a species and individuals, and allow assessing the population structure and the pattern of linkage disequilibrium. Resequencing a key germplasm subset (or core set) covering geographically well represented populations can contribute to build in-depth knowledge of genetic variation and diversity within a population based on sequence information (Sims et al. 2014). Additionally, sequencing the genomic DNA of a plant can generate a whole chloroplast genome sequence that has the potential to design a barcode for use in plant identification (Xu et al. 2012; Huang et al. 2010).

High-throughput resequencing has rapidly expanded our knowledge of genetic variations in crops as well as paved the way to develop molecular markers on a large scale. Coupled with high-throughput phenotyping, high-density molecular markers will help to accurately identify traits of interest, and eventually the markers are expected to increase breeding efficiency. Several types of molecular marker have been used for Brassica linkage analysis and molecular breeding purposes, including Random Amplified Polymorphic DNA(RAPD) (Tanhuanpää et al. 1995; Dos Santos et al. 1994; Hallden et al. 1994), Amplified Fragment Length Polymorphism(AFLP) (Pradhan et al. 2003; Negi et al. 2004), Sequence Characterized Amplified Region(SCAR) (Rahman et al. 2007), Simple Sequence Repeat (SSR) (Szewc-McFadden et al. 1996; Piquemal et al. 2005), and Sequence-related Amplified Polymorphism(SRAP) (Li and Quiros 2001). Recently SNPhas drawn more attention because SNPs are the most abundant class of polymorphisms found in plant and animal genomes (McNally et al. 2009; Feltus et al. 2004). Compared to SSR and other markers, SNP analysis can be done without requiring DNA separation by size using traditional gel method and therefore, can be automated in high throughput genotyping analysis. Moreover, biallelic nature of SNPs offers much lower error rate in

allele calling (Rafalski 2002). These advantages have resulted in SNPs increasingly becoming the markers of choice for accurate genotype identification and diversity analysis. Recently, high throughput SNP analysis in *Brassica* species including *B. rapa* (Park et al. 2010), *B. oleracea* (Izzah et al. 2014; Lee et al. 2015) and *B. napus* (Trick et al. 2009; Hayward et al. 2012) demonstrated the broad range of SNP applications (Bollina et al. 2015).

Polyploidization is a common mode of evolution in flowering plants, which occurs through genome hybridization (allopolyploidy) or genome duplication (autopolyploidy) and results in an increased gene set (Wang et al. 2012). The evolution of duplicated genes after whole-genome duplication (WGD) has been studied extensively. The loss/retention of duplicated genes is not a random process instead it appears to be depended on functional gene category in various plant species. In *A. thaliana*, over-retained genes are involved in basic cellular machinery, nucleotide-sugar metabolism, signal transduction or regulatory functions, while the diploidized genes are involved in DNA repair, tRNA ligation or defense (Blanc and Wolfe 2004). The loss/retention of duplicated genes might also depend on their parental origin. In this case, one of the parental genomes is more likely to retain genes and has a higher gene density than the other(s) genome(s). Whole genome duplication followed by structural and functional modifications may result in differential gene content or regulation in the duplicated regions, which can play a fundamental role in the diversification of genes underlying complex traits. *Brassica* is an outstanding model for investigating structural and/or functional comparisons of duplicated regions involved in the control of complex traits. *Brassica* ancestors have undergone two duplication events (named α and β) and two triplication events of which the most recent is specific to the *Brassica* clade (Chen and Birchler 2013). These WGD events along with hybridization of the two progenitor genomes have resulted in a large number of duplicated regions in the *Brassica*

genome. The ancestral Brassicaceae genome was reconstructed in 24 genomic blocks (A–X) also called the ancestral karyotype blocks (AK blocks). These blocks have been identified in *B. rapa* (Cheng et al. 2013), *B. oleracea* (Gulick et al. 2009), and *B. napus* (Parkin et al. 2014). Moreover, the structural organization of the hexaploid *Brassica* ancestor genome was determined from whole-genome sequence analyses of the *B. rapa* and *B. oleracea* genomes. Comparative genome analyses of *Brassica* regions involved in polygenic traits would then give the opportunity to study the impact of genome duplications on the structural and functional organization in a highly duplicated genome.

With an objective to understand the genome-wide diversity and evolution across *Brassica* species, a set 28 *Brassica* genotypes (Table 3) was re-sequenced up to 2-5X genome coverage on average using Illumina Mi-Seq platform with 300bp paired end mode. A total of 176 million paired reads were aligned to *A. thaliana*, *B. rapa*, *B. oleracea,* and *B. napus* reference genomes and, a total of 24 million high-confident genome-wide SNPs was discovered across the 28 genotypes using stringent variant calling strategies. The SNPs identified here may enhance the marker density of the existing genetic maps, which could also be a useful source for high-throughput QTL mapping and marker-assisted *Brassica* improvement. In addition, SNP based genome wide comparative analysis was performed among the seven *Brassica* genome groups to investigate the similarity/divergence. Altogether, these results not only provide tremendous opportunity to unravel the evolutionary history of *Brassica*, but they also serve as a valuable source to rapidly identify agronomically important genes in genus *Brassica*.

MATERIALS AND METHODS

Plant materials and whole genome resequencing

In order to investigate the genetic diversity and evolution of *Brassica* species in U's triangle, seven distinct genotype groups were selected and comparative analyses were conducted among the *Brassica* species (Table 3). Seeds representing four accessions of each of A, B, C, R, AB, AC, and BC genome were obtained from RDA-Genbank Center (http://www.genebank.go.kr/), Suwon, South Korea. All plant materials were prepared at 22/18℃ with 16/8 h light (night/dark) conditions in RDA experimental farm, Suwon, South Korea during spring season of 2014. Genomic DNA was extracted from approximately 5g of young leaves for all 28 genotypes following the modified cetyltrimethylammonium bromide (CTAB) protocol (Allen et al. 2006). Prior to Illumina library preparation, the quality and quantity of the DNA were examined using both PicoGreen assay and NanoDrop ND-1000 (NanoDropTechnologies, Inc., USA) (Table 3).

Table 3. List of *Brassica* accessions used in this study

Genome type	Sample Name	Sample Number	Accession Number	Genotypes	PicoGreen assay (ng/u	Purity (A260 A280)
A	A1	144	K194441	*Brassica rapa suB. rapa*	58	2.08
	A2	156	K201443	*Brassica rapa subsp. chinensis*	158	1.87
	A3	192	K128851	*Brassica rapa subsp. pekinensis*	181	1.84
	A4	211	137572	*Brassica rapa sub. dichotoma*	176	1.84
B	B1	106	119404	*Brassica nigra*	193	1.74
	B2	133	135140	*Brassica nigra*	176	1.85
	B3	136	135242	*Brassica nigra*	166	1.83
	B4	95	119326	*Brassica nigra*	210	1.76

C	C1	H1	-	*Brassica oleracea var.capitata*	164	1.83
	C2	H17	-	*Brassica oleracea var.*	177	1.83
	C3	H20	-	*Brassica oleracea var.*	196	1.83
	C4	H43	-	*Brassica oleracea var.*	188	1.83
R	R1	290	K046542	*Raphanus raphanistrum*	165	1.81
	R2	294	K036707	*Raphanus sativus var. raphanistroides*	196	2.1
	R3	319	100594	*Raphanus sativus var. sativus*	232	2.09
	R4	363	K012402	*Raphanus sativus var. sativus*	179	1.81
AB	AB1	189	K201621	*Brassica juncea var. integrifolia*	166	1.81
	AB2	231	K139720	*Brassica juncea var. integrifolia*	175	1.68
	AB3	252	K201580	*Brassica juncea var*	122	1.81
	AB4	32	K128855	*Brassica juncea var. integrifolia*	152	1.65
AC	AC1	17	135171	*Brassica napus var. napus*	207	1.66
	AC2	90	134598	*Brassica napus var. napoBrassica*	214	1.72
	AC3	91	134599	*Brassica napus var. napoBrassica*	153	1.82
	AC4	92	134604	*Brassica napus var. napus*	168	1.85
BC	BC1	11	119514	*Brassica carinata*	198	1.61
	BC2	16	135088	*Brassica carinata*	202	1.72
	BC3	19	135246	*Brassica carinata*	161	1.84
	BC4	9	119511	*Brassica carinata*	145	1.82

More than 5 μg of extracted DNA was randomly sheared and quantified using DNA 1000 kit (Agilent Technologies, Inc., USA). A genomic DNA library was constructed using a multiplexed paired-end DNA sample prep kit with the manufacturer's protocols (Illumina Inc., USA). Purity and yield of the libraries were confirmed using the 2100 Bioanalyzer (Agilent Technologies, Santa Clara, USA) and the libraries were pooled and sequenced on Illumina Mi-seq (2 x 300bp) up to 2-5x genome coverage. Library construction and sequencing were carried out at the LABGENOMICS Company (Seoul, Korea). Sequence reads of each accession were de-convoluted using index sequences.

Post-sequencing analysis and SNP discovery

Overall processes of SNP discovery are described in Figure 2. Briefly, low-quality bases in raw reads (2 X 300 bp) were removed using Trimmomatic (Bolger et al. 2014) and the trimmed paired reads of each accession were aligned to the reference genome sequence of *A. thaliana*, *B. rapa*, *B. oleracea* and *B. napus* using BWA0.7.10 program (http://bio-bwa.sourceforge.net). Read grouping and removal of PCR duplicates were carried out using Picard 1.112 (http://picard.sourceforge.net) and misalignments caused by indels were corrected by local re-alignment using Genome Analysis Toolkit (GATK). The candidate SNPs were called using Variant Caller, a utility in GATK 3.1 (Bauer 2011). All the tools mentioned above were used with default parameters. To avoid false positives variants, candidate SNPs exhibiting any of the following conditions were removed: (1) mapping quality score lower than 4; (2) quality less than 30; and (3) less than 10x or more than 45x mapping depth. SNPs identified and filtered in each accession were merged and compared to each other to identify promising SNPs.

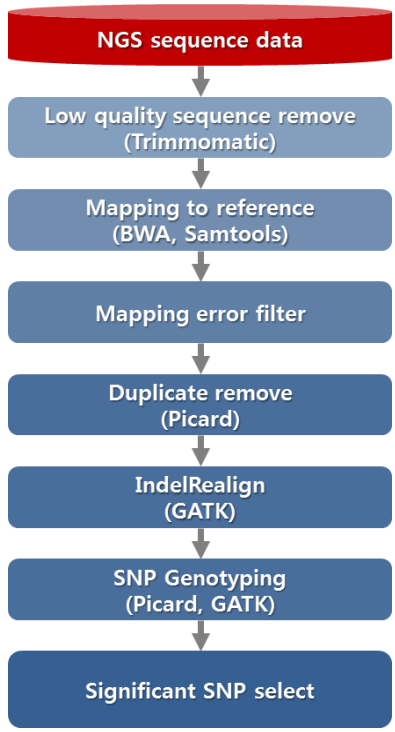

Figure 2. Schematic representation of the SNP discovery pipeline used to identify, validate and analyze putative SNPs in *Brassica* genomes using whole genome resequencing data.

Genome wide SNP based phylogenetictree construction

In order to identify the evolutionary relationship among the *Brassica* accessions, the filtered vcf files were converted to hapmap file and SNPhylo program (Lee et al. 2014) was used to filter SNPs based on MAF (minor allele frequency), missing rate, and linkage equilibrium then the remaining SNPs were concatenated to make a sequence file. Phylogenetic tree is constructed with Maximum likelihood method using MEGA 7 tool (Hall 2013).

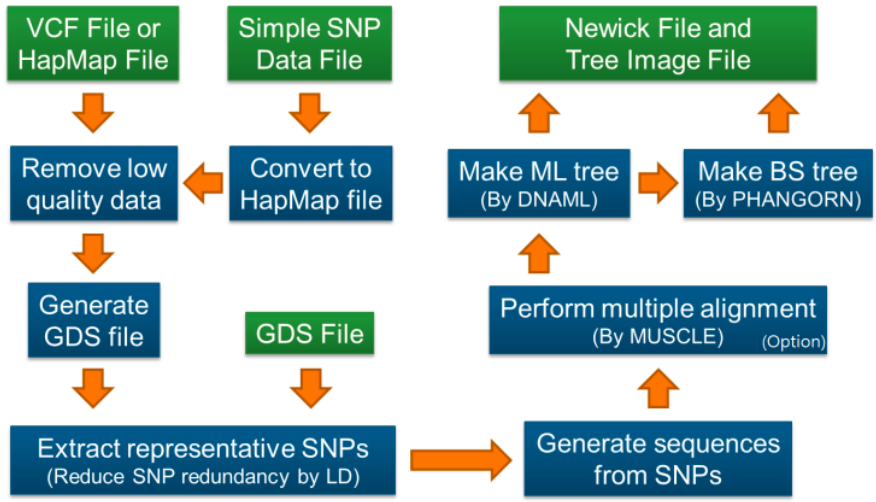

Figure 3. Schematic representation of the phylogenetic tree building using SNPhylo tool(Lee et al. 2014).

Besides, SNP based kinship analysis was performed using Genome Association and Prediction Integrated Tool (GAPIT) package (Lipka et al. 2012) with Efficient Mixed Model Association (EMMA) algorithmutilizing vcf file of the 28 *Brassica* genus mapped to *A. thaliana*, *B. rapa*, *B. napus* and *B. oleracea* genome which was completed genome sequencing. In order to examine kinship patterns depending on the number of SNP in same position; firstly, VCF file for genomes of *B. rapa*, *B. napus* and *B. oleracea* were applied to three groups: 5, 10 and 15 SNP. Then, kinship analysis was performed to retrieve genotyping data (call rate=0.8).

RESULTS

Whole genome resequencing and mapping of 28 *Brassica* genotypes

Whole genome of 28 *Brassica* accessions was resequenced using Illumina MiSeq next-generation sequencing platform. Approximately 194 million raw reads (about 58.3 Gb in total) were generated and 80.3 % of sequences were considered a high quality base (Q30 or higher). About 86% of the raw reads were passed trimming conditions and the remaining number of reads were ranged 4.7~8.5 million sequences (Table 4) per accession. Of the total pre-processed sequence reads (176 million reads), 22-54% were properly mapped to unique position in reference genome of *A. thaliana* (125 Mb), *B. rapa* (529 Mb), *B. oleracea* (696 Mb), and *B. napus* (849 Mb) with default mapping parameters (Table 5). The average mapping depth was 3-4 folds for each accession and the filtered reads were uniformly distributed across the chromosomes of each reference genome.

As expected, the rate of properly and uniquely mapped reads seemed to depend on the reference genome type. When A genome type used as a reference (*B. rapa*), A1-A4 accessions showed the highest proper mapping rate (54.8% on average) followed by AC1-AC4 (54.1%) and AB1-AB4 accessions (51.6%). With *B. oleracea* (C genome type) as a reference, 62.9% of C1-C4 reads were properly mapped followed by AC1-AC4 (58.7%) and A1-A4 accessions (56%). Interestingly, *Raphanus* accessions (R1-R4) were properly mapped to A and C genome references with 49% and 51.4%, respectively, and the proper mapping rate was higher than that of *B. nigra* (B genome type) accessions at 41.4% to A genome reference and 41.6% to C genome reference. Like previous reports pointed out (Wang 2013, Huang et al. 2016), the results also suggested that *B. rapa* and *B. oleracea* are evolutionary closer to *Raphanus* genome than *B. nigra.* This was confirmed again when 28 accessions were mapped to *B. napus* (AC

genome type) reference genome. As we expected, C1-C4 and AC1-AC4 showed the highest proper mapping rates (63.4% and 63.3%, respectively) followed by A accessions (59%). R genome accessions showed 52.9% of proper mapping rate on average, which was much higher than that of B and BC genome accessions (42.2% and 46.5%, respectively).

Brassica species were diverged from *Arabidopsis* around 17 MYA therefore *Arabidopsis* can be the best reference to compare all 28 accessions. However the results of proper mapping rate were different from what was observed above. In diploid accessions (A, B, C, R), *Raphanus* (R type) showed the highest proper mapping rate (26.5%) followed by A and C (22.9% and 22.7% respectively), and B genome type (*B. nigra*) was the least (20.5%). Further investigation is needed on these results, but *B. nigra* seemed to experience faster evolutionary forces than *Raphanus* species after diverging from *Arabidopsis*.

SNPs and indels across 28 *Brassica* species

In order to assess sequence variations from 28 *Brassica* accessions, high-quality reads were aligned to *A. thaliana*, *B. rapa*, *B. oleracea*, and *B. napus* reference genomes, and called SNPs using the BWA and SAMtools programs. To detect homozygous polymorphisms, two filters were applied: a minimum of four reads and a maximum of 128 reads had to be mapped at any position and a minimum allele frequency of 0.9 was required. When the minimal read depth was increased to eight, the number of SNPs was dramatically decreased and several polymorphic SNPs previously determined by Sanger sequencing were no longer detected. Therefore, a minimum of four reads is an optimal to filter spurious SNPs and reduce false positive SNPs. Using the above, stringent variant calling strategies, a total of 2.8, 22.6, 20.9 and 13.1 million high-confidence, genome-wide SNPs were detected with references of *A. thaliana*, *B. rapa*, *B. oleracea* and

B. napus, respectively (Table 6). The frequency of SNPs was varied depending on the reference used and was 193, 29, 43, and 83 bp per SNP with *A. thaliana*, *B. rapa*, *B. oleracea*, and *B. napus*, respectively. The total number of SNPs detected varied widely from one species as a reference to another, with a range of 10~15%. However, C genome type (C1-C4) species showed comparatively higher SNPs in all four reference genomes. The total number of SNPs also varied widely between the different chromosomes in all four reference genomes. The range of variation between the chromosomes reached 10-folds on average from the reference. *B. rapa* and *B. oleracea* showed high variation of SNPs compared to other two references.

In addition, 8,688; 95,669; 83,434; and 59,616 unique indels were detected when 28 accessions were separately mapped to *A. thaliana*, *B. rapa*, *B. oleracea*, and *B. napus*, respectively, with the default parameters (Table 6). This number varied from 287 to 8,251 in *B. rapa* and from 116 to 193 in *A. thaliana* as a reference. Their distributions across 28 species were more homogeneous than that of SNPs, although *B. rapa*, with a high density compared to the average, could be detected. In most cases, the *Brassica* species carrying a high number of SNPs also exhibited a high number of indels. The correlation between SNP and indel numbers on each species was higher than 0.98. The majority of indels corresponded to a unique base modification, but a maximum of 32 bp deletions and 25 bp insertions were also detected. The number of deletions was a little higher than the number of insertions (with a ratio varying from 1.09 to 1.44) according to the species and respective reference genome.

In order to find the association between SNPs and genome annotations, SNPs found in exons, introns, downstream, upstream, and intergenic regions were examined. As a result, the similar SNP distribution was shown in downstream and upstream regions of all reference genomes. Whereas the SNP frequency detected in exon, intron and intergenic regions

were different. Most of the SNPs were distributed in exon regions in *A. thaliana*, while others showed that higher number of SNPs distributed in intergenic region. The highest polymorphisms (32.4%) were detected in upstream region (Figure 4).

Table 4. Raw read and pre-processing summary of whole genome resequencing data of 28 *Brassica* accessions.

Group	sample name1	Sample name2	Yield (bp)	Reads (#)	Trimmed reads (#)	% paired after trim	GC(%)	Q20(%)	Q30(%)
A	A1	cabagge144	2,061,447,577	6,868,530	5,997,255	87.31%	39.72	90.23	80.9
	A2	cabbage156	1,633,838,458	5,445,034	4,720,368	86.69%	40.24	89.61	80.08
	A3	cabbage192	1,824,582,285	6,077,818	5,069,866	83.42%	40.87	88.49	78.28
	A4	cabbage211	2,020,800,562	6,732,026	5,603,598	83.24%	41.18	88.1	77.79
B	B1	cabbage106	2,112,950,293	7,035,680	5,875,186	87.51%	40.29	91.17	82.01
	B2	cabbage133	1,983,078,477	6,601,940	6,200,635	88.13%	40.36	90.52	80.56
	B3	cabbage136	2,121,350,728	7,063,090	5,706,132	86.43%	40.01	91.25	81.77
	B4	cabbage95	2,013,201,544	6,713,744	6,191,161	87.66%	42.16	90.78	81.1
C	C1	hwaseo1	1,959,036,582	6,539,790	5,709,832	87.31%	42.8	90.67	81.36
	C2	hwaseo7	1,733,587,435	5,812,614	5,051,563	86.91%	39.91	90.28	81.56
	C3	hwaseo20	2,134,675,000	7,147,450	6,241,028	87.32%	39.82	90.45	81.29
	C4	hwaseo43	2,677,165,673	8,918,982	8,036,206	90.10%	39.51	91.91	83.97
AB	AB1	cabbage189	1,776,503,201	5,921,522	5,710,182	84.95%	39.91	90.78	81.4
	AB2	cabbage231	2,003,442,508	6,700,220	5,185,766	87.57%	40.07	90.52	82.03
	AB3	cabbage252	2,027,224,339	6,792,186	5,888,063	87.88%	40.62	90.71	82.2
	AB4	cabbage32	2,016,577,985	6,721,494	5,982,831	88.08%	40.33	89.35	79.23
	AC1	cabbage17	3,171,142,796	10,567,346	8,528,155	80.70%	41.23	87.51	76.31

AC	AC2	cabbage90	2,064,273,213	6,887,546	5,673,257	82.37%	42.41	88.55	77.98
	AC3	cabbage91	2,231,986,508	7,435,044	6,563,185	88.27%	39.49	90.82	81.69
	AC4	cabbage92	2,002,591,113	6,671,374	5,828,281	87.36%	40.57	90.23	80.89
BC	BC1	cabbage11	2,007,552,783	6,691,520	5,993,222	85.32%	40.45	88.24	76.98
	BC2	cabbage16	2,470,907,980	8,237,636	5,549,020	82.93%	40.31	89.41	79.02
	BC3	cabbage19	2,025,099,949	6,750,460	5,549,020	82.93%	39.84	91.54	82.55
	BC4	cabbage9	2,107,824,688	7,024,784	6,010,547	89.04%	40.48	89.31	78.92
R	R1	cabbage290	1,983,648,676	6,659,950	5,731,710	86.06%	40.21	89.63	80.46
	R2	cabbage294	1,975,605,296	6,611,620	5,780,862	87.43%	39.65	90.26	81.52
	R3	cabbage319	2,045,480,101	6,859,446	5,651,074	82.38%	42.78	88.72	78.2
	R4	cabbage363	2,086,711,877	6,967,708	5,732,693	82.28%	41.22	88.98	78.39

Table 5. Summary of mapping result with four Brassicaceae family genomes.

genome	sample name1	Sample name2	Raw Reads	*B. oleracea* (C)				*B. rapa* (A)				*B. napus* (AC)				*Arabidopsis*			
				*Mapped	Properly Paired	% Properly paired	AVG Depth	Mapped	Properly Paired	% Properly paired	AVG Depth	Mapped	Properly Paired	% Properly paired	AVG Depth	Mapped	Properly Paired	% Properly paired	AVG Depth
A	A1	cabbage144	6,266,688	7,957,267	3,486,710	55.6%	4.59	7,146,039	3,488,232	55.7%	3.76	7,563,746	3,772,718	60.2%	3.28	2,036,823	1,483,488	23.7%	4.59
A	A2	cabbage156	4,950,806	6,258,123	2,701,982	54.6%	3.97	5,664,345	2,673,652	54.0%	3.1	6,067,665	2,814,796	56.9%	2.78	1,540,498	1,129,188	22.8%	4.11
A	A3	cabbage192	5,484,758	6,808,932	3,135,630	57.2%	4.41	6,203,362	3,113,186	56.8%	3.41	6,653,106	3,280,618	59.8%	2.99	1,729,533	1,270,592	23.2%	3.84
A	A4	cabbage211	6,045,542	7,344,161	3,412,468	56.4%	4.76	6,762,746	3,186,330	52.7%	3.55	7,170,027	3,581,164	59.2%	3.11	1,862,612	1,338,556	22.1%	4.48
B	B1	cabbage106	6,406,822	6,385,909	2,581,128	40.3%	3.79	5,448,742	2,622,022	40.9%	3.16	6,812,012	2,604,408	40.7%	2.97	1,654,206	1,254,942	19.6%	3.4
B	B2	cabbage133	5,932,546	5,789,791	2,399,702	40.4%	3.78	4,927,669	2,342,336	39.5%	3.03	6,144,979	2,448,030	41.3%	2.81	1,516,383	1,130,400	19.1%	3.37
B	B3	cabbage136	6,366,880	6,412,781	2,594,982	40.8%	3.72	5,486,340	2,653,396	41.7%	3.17	6,862,271	2,603,006	40.9%	3.02	1,677,448	1,281,956	20.1%	3.55
B	B4	cabbage95	6,092,306	6,139,981	2,734,814	44.9%	4.42	5,222,483	2,656,678	43.6%	3.49	6,454,751	2,795,514	45.9%	3.23	1,849,388	1,417,314	23.3%	3.91
C	C1	hwaseo1	6,003,462	6,455,486	4,007,438	66.8%	3.74	7,197,155	3,222,508	53.7%	3.88	6,888,971	4,001,798	66.7%	3.29	1,950,958	1,534,144	25.6%	4.6
C	C2	hwaseo17	5,309,784	5,764,383	3,218,818	60.6%	3.01	6,388,427	2,499,896	47.1%	3.02	6,091,606	3,330,288	62.7%	2.57	1,568,156	1,150,438	21.7%	4.63
C	C3	hwaseo20	6,595,034	7,142,780	4,051,456	61.4%	3.34	8,043,806	3,241,522	49.2%	3.37	7,607,746	4,048,802	61.4%	2.87	1,897,575	1,454,016	22.0%	4.2
C	C4	hwaseo43	8,451,840	9,334,665	5,296,786	62.7%	4.08	10,527,504	4,068,652	48.1%	4.04	9,958,899	5,325,282	63.0%	3.47	2,422,888	1,826,272	21.6%	5.08
R	R1	cabbage290	6,066,826	5,465,151	2,815,518	46.4%	4.78	4,915,009	2,723,142	44.9%	3.5	5,844,915	2,894,626	47.7%	4.02	1,814,330	1,362,206	22.5%	3.63
R	R2	cabbage294	6,145,984	5,860,767	3,142,318	51.1%	5.46	5,283,069	2,939,032	47.8%	3.82	6,163,790	3,253,492	52.9%	4.22	2,029,696	1,542,014	25.1%	3.82
R	R3	cabbage319	6,069,312	5,666,085	3,181,612	52.4%	5.83	5,033,130	3,012,642	49.6%	4.21	5,870,388	3,294,224	54.3%	4.76	2,187,247	1,711,074	28.2%	4.44
R	R4	cabbage363	6,113,012	6,108,515	3,389,830	55.5%	5.78	5,439,220	3,269,566	53.5%	4.21	6,364,960	3,473,968	56.8%	4.87	2,324,687	1,838,476	30.1%	4.66
AB	AB1	cabbage189	5,399,244	6,111,909	2,610,884	48.4%	3.17	5,412,116	2,889,670	53.5%	2.56	6,092,327	2,677,006	49.6%	2.21	1,541,563	1,190,058	22.0%	3.35

AB	AB2	cabbage231	6,195,310	6,799,048	2,935,474	47.4%	3.47	5,982,882	3,176,512	51.3%	2.72	6,784,403	3,026,234	48.8%	2.32	1,718,824	1,280,552	20.7%	3.53
AB	AB3	cabbage252	6,327,028	6,947,097	2,954,692	46.7%	3.42	6,227,424	3,302,336	52.2%	2.79	7,006,762	3,074,388	48.6%	2.39	1,734,109	1,311,654	20.7%	3.3
AB	AB4	cabbage32	6,000,246	6,538,859	2,772,358	46.2%	3.45	5,819,552	2,969,688	49.5%	2.66	6,604,843	2,928,964	48.8%	2.25	1,573,775	1,128,536	18.8%	3.47
AC	AC1	cabbage17	9,217,472	10,405,825	5,255,564	57.0%	3.74	10,702,275	4,873,944	52.9%	4.04	10,553,086	5,655,556	61.4%	2.67	2,497,445	1,875,004	20.3%	4.55
AC	AC2	cabbage90	6,027,702	6,877,916	3,741,840	62.1%	3.54	6,996,320	3,439,566	57.1%	3.61	6,958,554	3,992,496	66.2%	2.44	2,089,570	1,559,836	25.9%	4.36
AC	AC3	cabbage91	6,859,838	8,026,562	4,011,946	58.5%	3.28	8,214,628	3,607,540	52.6%	3.42	7,978,001	4,389,426	64.0%	2.32	1,923,735	1,380,912	20.1%	4.18
AC	AC4	cabbage92	6,128,616	7,070,720	3,512,222	57.3%	3.01	7,304,498	3,311,592	54.0%	3.18	7,110,180	3,765,252	61.4%	2.18	1,717,395	1,283,878	20.9%	3.54
BC	BC1	cabbage11	5,913,424	5,765,694	2,843,604	48.1%	2.4	5,799,074	2,517,026	42.6%	2.55	6,060,106	2,761,846	46.7%	1.98	1,477,577	1,112,864	18.8%	3.18
BC	BC2	cabbage16	7,401,876	7,498,717	3,697,846	50.0%	2.79	7,562,908	3,282,816	44.4%	2.96	7,893,404	3,565,130	48.2%	2.28	1,917,149	1,447,908	19.6%	3.63
BC	BC3	cabbage19	6,224,152	6,293,789	2,959,292	47.5%	2.45	6,420,470	2,649,550	42.6%	2.6	6,627,603	2,708,406	43.5%	2.08	1,467,031	1,096,948	17.6%	3.19
BC	BC4	cabbage9	6,324,410	6,246,237	3,071,904	48.6%	2.54	6,317,452	2,715,064	42.9%	2.71	6,543,421	3,001,240	47.5%	2.07	1,678,600	1,256,514	19.9%	3.43

Table 6. SNP variants identified in four 28 *Brassica* genomes

Sample name	*B. rapa*					*B. napus*					*B. oleracea*					*A. thaliana*				
	Variation rate	Number of Variants (by type)			Total Variants	Variation rate	Number of Variants (by type)			Total Variants	Variation rate	Number of Variants (by type)			Total Variants	Variation rate	Number of Variants (by type)			Total Variants
		SNP	INS	DEL			SNP	INS	DEL			SNP	INS	DEL			SNP	INS	DEL	
A1	293	867,018	3,753	4,329	875,393	974	656,798	2,631	2,727	662,156	294	1,296,918	5,540	5,741	1,308,199	1,008	117,767	132	193	118,092
A2	447	571,893	1,609	1,824	575,326	1,166	550,064	1,589	1,688	553,341	395	968,451	2,381	2,461	973,293	1,354	87,647	129	182	87,958
A3	530	481,887	1,623	1,707	485,217	1,072	597,420	2,066	2,263	601,749	371	1,031,560	3,033	3,089	1,037,682	1,224	97,023	123	190	97,336
A4	336	757,694	2,964	3,337	763,995	993	644,988	2,330	2,453	649,771	368	1,039,434	3,261	3,253	1,045,948	1,232	96,321	128	183	96,632
B1	455	563,767	452	446	564,665	3,011	213,847	217	219	214,283	786	488,613	425	440	489,478	1,043	113,934	122	177	114,233
B2	536	479,284	287	276	479,847	3,572	180,370	129	150	180,649	921	417,381	280	303	417,964	1,208	98,297	116	184	98,597
B3	444	577,711	472	478	578,661	2,926	220,076	222	223	220,521	766	501,451	459	468	502,378	1,024	115,955	128	186	116,269
B4	481	533,216	364	396	533,976	3,151	204,412	177	216	204,805	827	464,663	390	408	465,461	1,055	112,605	126	185	112,916
C1	224	1,141,251	2,942	3,349	1,147,542	859	746,249	2,231	2,362	750,842	776	493,346	1,338	1,333	496,017	1,154	102,883	127	182	103,192
C2	251	1,018,123	1,973	2,108	1,022,204	965	665,520	1,490	1,578	668,588	573	667,703	1,572	1,765	671,040	1,591	74,545	129	179	74,853
C3	203	1,259,018	3,477	3,584	1,266,079	797	805,168	2,073	2,101	809,342	510	749,830	1,926	1,890	753,646	1,287	92,202	129	180	92,511
C4	155	1,636,282	7,217	8,251	1,651,750	560	1,140,478	5,035	5,494	1,151,007	369	1,033,276	4,665	5,270	1,043,211	997	119,171	135	186	119,492
R1	529	484,986	326	306	485,618	3,778	170,522	146	127	170,795	936	410,376	324	292	410,992	1,269	93,556	134	180	93,870
R2	555	462,717	318	286	463,321	3,896	165,315	179	162	165,656	975	394,102	332	325	394,759	1,328	89,346	144	188	89,678
R3	554	463,608	289	273	464,170	3,787	170,059	183	138	170,380	969	396,452	357	309	397,118	1,265	93,848	128	177	94,153
R4	517	496,503	421	360	497,284	3,531	182,321	237	197	182,755	910	422,192	432	372	422,996	1,193	99,515	135	181	99,831
AB1	429	597,693	570	643	598,906	1,592	404,200	452	495	405,147	541	709,288	665	587	710,540	1,232	96,392	120	180	96,692
AB2	363	706,206	839	912	707,957	1,390	463,091	593	549	464,233	468	820,504	769	774	822,047	1,094	108,572	122	181	108,875
AB3	344	744,170	941	1,088	746,199	1,319	487,812	680	706	489,198	453	847,695	815	840	849,350	1,077	110,314	124	183	110,621
AB4	390	658,232	679	737	659,648	1,519	423,824	494	488	424,806	512	750,529	655	657	751,841	1,201	98,879	122	180	99,181
AC1	172	1,477,857	4,448	4,963	1,487,268	1,038	618,459	1,467	1,513	621,439	274	1,393,433	3,495	3,814	1,400,742	894	132,864	136	187	133,187
AC2	265	964,987	2,180	2,444	969,611	1,495	429,976	850	865	431,691	410	933,967	1,844	1,947	937,758	1,094	108,510	130	185	108,825
AC3	200	1,280,029	2,915	3,151	1,286,095	1,166	551,107	1,058	1,051	553,216	332	1,151,756	2,120	2,313	1,156,189	1,057	112,312	122	182	112,616
AC4	244	1,047,057	1,549	1,685	1,050,291	1,616	398,218	487	522	399,227	400	958,174	1,177	1,313	960,664	1,159	102,414	125	181	102,720

BC1	365	702,534	515	467	703,516	1,539	418,606	347	339	419,292	723	531,285	411	428	532,124	1,497	79,234	119	186	79,539
BC2	248	1,033,506	1,106	1,155	1,035,767	1,028	625,744	782	837	627,363	488	786,126	938	980	788,044	1,053	112,788	125	192	113,105
BC3	292	878,044	773	744	879,561	1,223	526,420	576	549	527,545	568	675,653	643	687	676,983	1,249	95,074	121	187	95,382
BC4	323	793,184	681	687	794,552	1,357	474,722	414	469	475,605	638	601,475	547	581	602,603	1,339	88,654	121	179	88,954

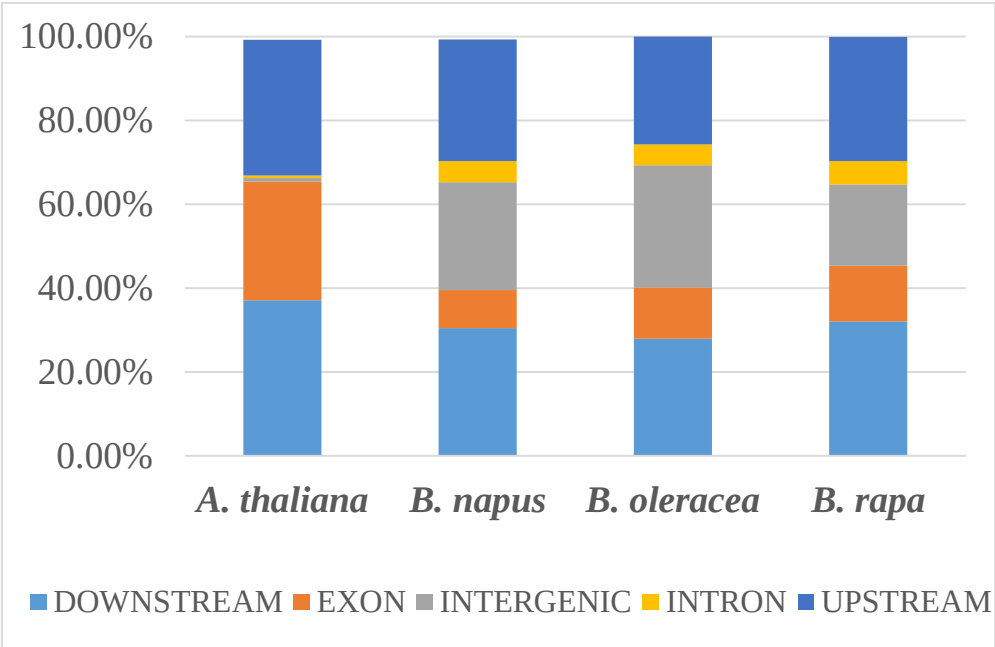

Figure 4. Distribution of SNPs from 28 *Brassica* genotypes across the genomic regions in *A. thaliana, B. napus, B. oleracea, and B. rapa*.

Genome wide SNP based phylogenetic tree construction

Phylogenetic trees are widely used for genetic and evolutionary studies in various organisms. Advanced sequencing technology has dramatically enriched data available for constructing phylogenetic trees based on SNPs. In order to investigate the genetic relationship among the different *Brassica* species, a maximum likelihood phylogenetic tree was constructed using the SNP data (Figure 5) with SNPhylo and Mega7. This phylogenetic tree indicated that the divergence between the *Arabidopsis* and *Brassica* species, which is consistent with other reports (Park et al. 2005; Li et al. 2005).

Diploid genome positions (A, B, C, R) in phylogenetic trees generated using SNPs with *B. rapa*, *B. oleracea*, and *B. napus* references agreed with each other, of which B genome was first diverged from *Arabidopsis* then *Raphanus*, and finally A and C genome, (B(R(C,A))). Although genetic distances in three different trees were varied based on which genome was used for mapping, this topology is, what we believe at present, interpret *Brassica* speciation. However, with *Arabidopsis* as a reference, tree topology appears that B and R were separated from A and C. When tetraploid genomes (AB, AC, BC) were added to construct phylogenetic trees, three genomes were localized in different positions. In phylogenetic trees based on SNPs from *B. napus* (AC genome type) and *B. oleracea* (C genome type), all of the species positions were congruent to each other (Figure 5 C and D). However, with *B. rapa* as reference, all three tetraploid (AC, AB, BC) were differently located compared with the above results (Figure 5. B), thus further analyses of simple genome such as CpDNA and nr DNA were required to understand tetraploid genomes in *Brassica* speciation.

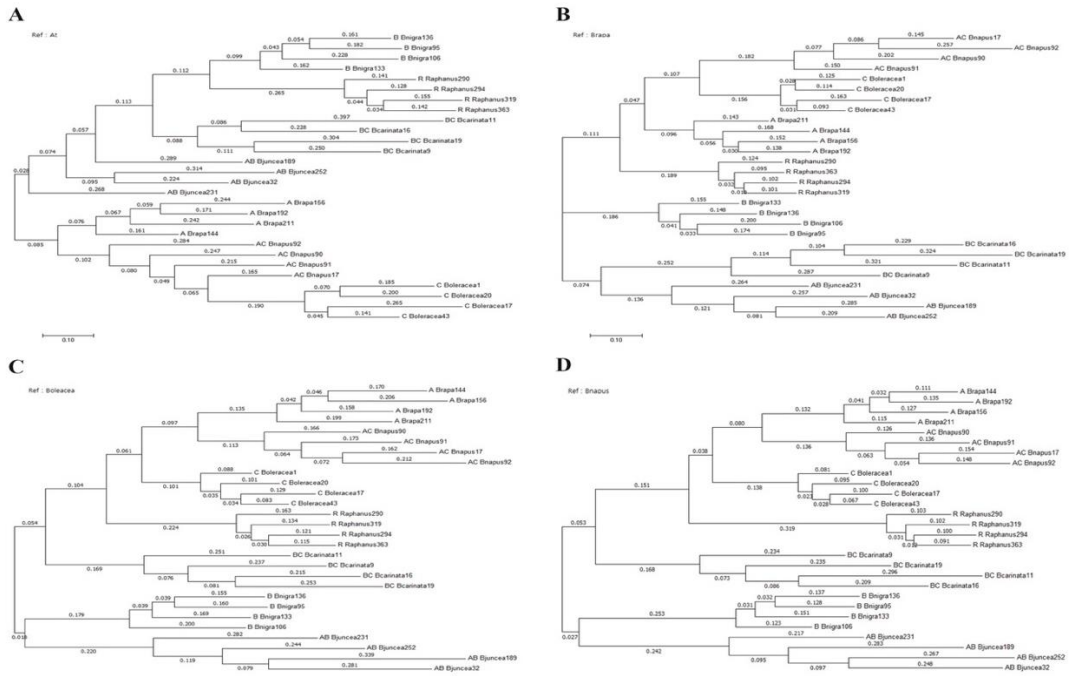

Figure 5. Phylogenetic tree analysis of *Brassica* accessions. (A) *A. thaliana* (B) *B. rapa* (C) *B. oleracea* and (D) *B. napus*. A maximum-likelihood tree was constructed with SNP data from each genomes using SNPhylo program and visualized using MEGA7 tool.

With *B. oleracea* as reference genome, GAPIT program was performed to conduct the kinship analysis with VCP file to diploid 16 accessions. Table 7 shows the number of SNP genotyping for 5, 10, and 15 accessions among SNPs to meet the requirement of read mapping depth and the number of SNP in the total vcf files. This analysis found that the same result was shown when *B. rapa* genome was used as a reference genome. It was examined that there was clear distinction between A, B, C, and R genome under all conditions, and similar species have close relationship with one another. On the other hand, as the study describes Figure 6, the tree of Figure 6 is the complete opposite from their actual evolution process.

Using *B. oleracea* genome as a reference, four accessions of C genome would be the most mapped compared to the other accessions. As the genome divergence time is closer, genotyping number of SNP will be less. Although the number of SNP is reduced a lot, it is appropriate to be analyzed using commonly genotyped SNP in 15 accession more. However, Figure 8 does not show the same tree of previously reported evolution studies of diploid *Brassica* genus.

Table 7. SNP numbers in *B. oleracea* genome mapping result for kinship analysis

# of progenies	C01	C02	C03	C04	C05	C06	C07	C08	C09	Total
vcf file	908,100	933,328	1,317,888	911,339	752,343	875,114	1,121,621	953,547	925,379	8,698,659
5	420,770	348,399	575,286	403,844	351,376	372,625	480,441	438,158	413,660	3,804,559
10	116,699	76,679	157,246	109,924	101,512	109,512	133,538	123,894	112,999	1,042,003
15	9,372	5,038	12,474	7,703	7,758	10,186	10,435	10,300	8,426	81,692

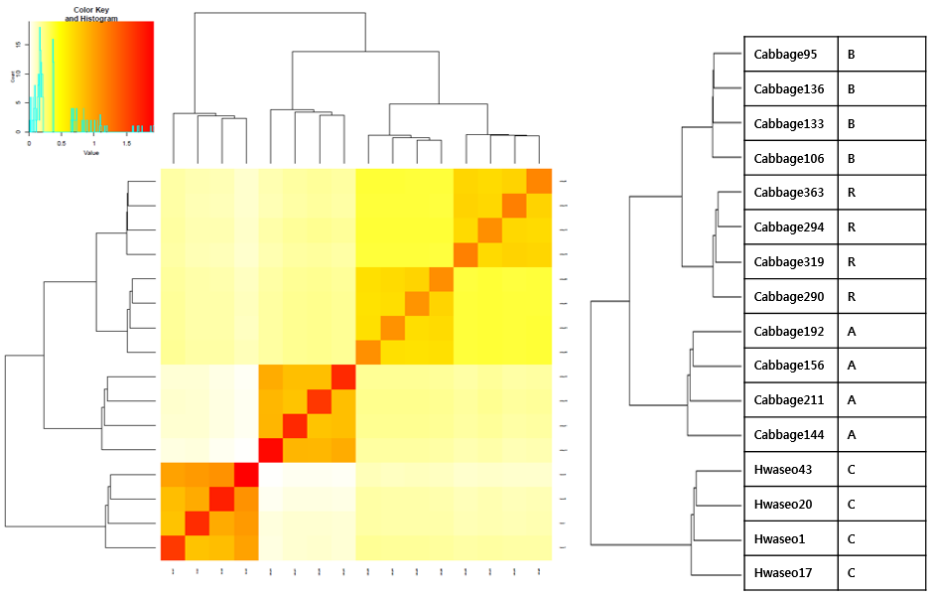

Figure 6. Heat map of a kinship matrix of diploid *Brassica* accessions based on 5 more SNPs genotyping in *B. oleracea*

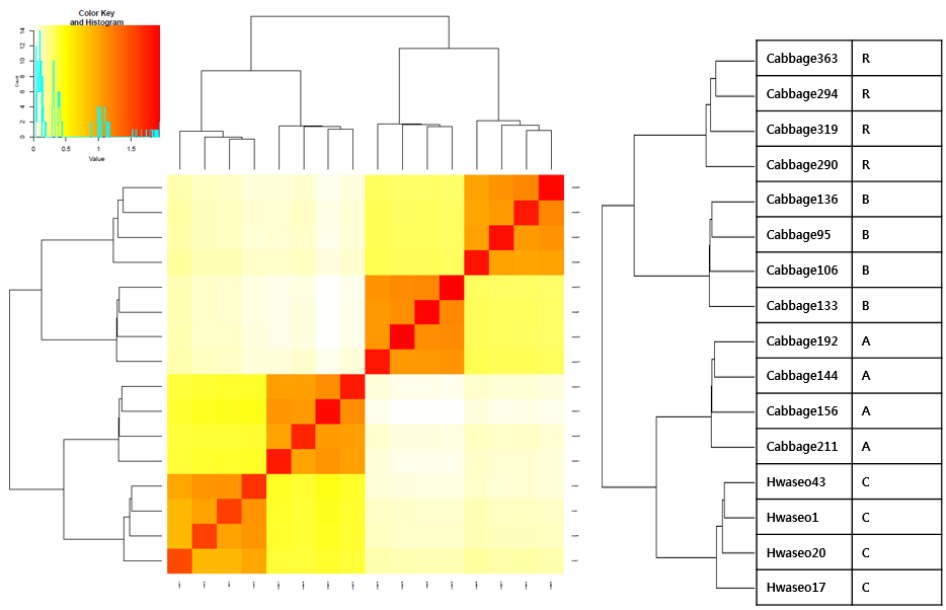

Figure 7. Heat map of a kinship matrix of diploid *Brassica* accessions based on 10 more SNPs genotyping in *B. oleracea*

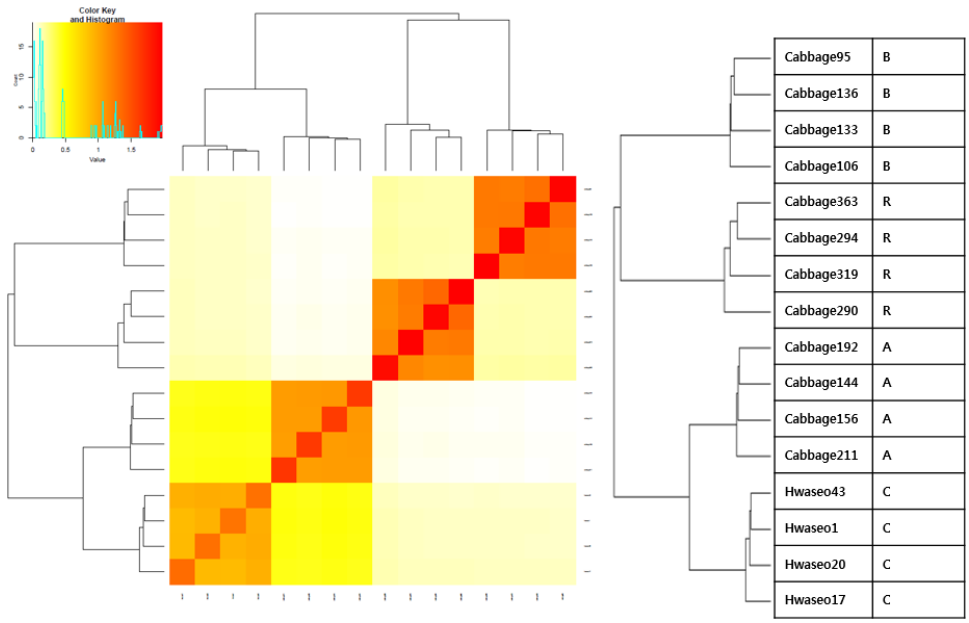

Figure 8. Heat map of a kinship matrix of diploid *Brassica* accessions based on 15 more SNPs genotyping in *B. oleracea*

Overall, this study confirmed that it was appropriate to analyze SNPs which has many accessions of genotyping. Bias from reference genome could be more efficiently removed by analyzing *A. thaliana* genome as reference, rather than using *B. rapa* or *B. olearacea* genome. Thus, GAPIT program was performed by using VCF file with SNP calling to diploid 16 accessions to apply kinship analysis as *A. thaliana* reference genome. Table 8 shows the number of selected SNP and its result is described in Figure 9. It was observed that there was significant difference between each diploid A, B, C, and R genome individually, and their relationship matched with their genomes' evolutionary process.

Table 8. SNP numbers of diploid *Brassica* accessions based on 15 more SNPs genotyping in *A. thaliana*

# of progenies	chr01	chr02	chr03	chr04	chr05	total
vcf file	150,814	85,696	115,948	89,837	141,929	584,224
15	341	1,711	455	240	353	3,100

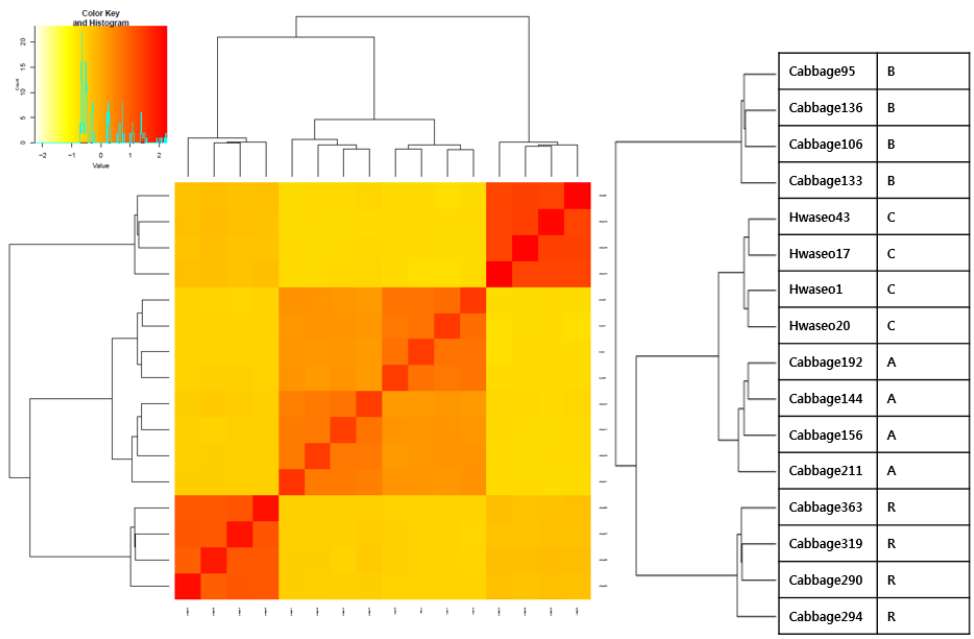

Figure 9. Heat map of a kinship matrix of diploid *Brassica* accessions based on 15 more SNPs genotyping in *A. thaliana*

This study confirmed that it was appropriate to analyze SNPs which has many accessions of genotyping using *A. thaliana* as reference genome. A total 28 *Brassica* accessions were analyzed in the same manner and this work screened SNP having more than 25 accessions with genotyping among 28 accessions. The number of SNP and their result is shown in Table 9 and Figure 10, respectively.

Table 9. SNP numbers of 28 *Brassica* accessions based on 25 more SNPs genotyping in *A. thaliana*

# of progenies	chr01	chr02	chr03	chr04	chr05	total
vcf file	150,814	85,696	115,948	89,837	141,929	584,224
25	534	1844	644	384	506	3912

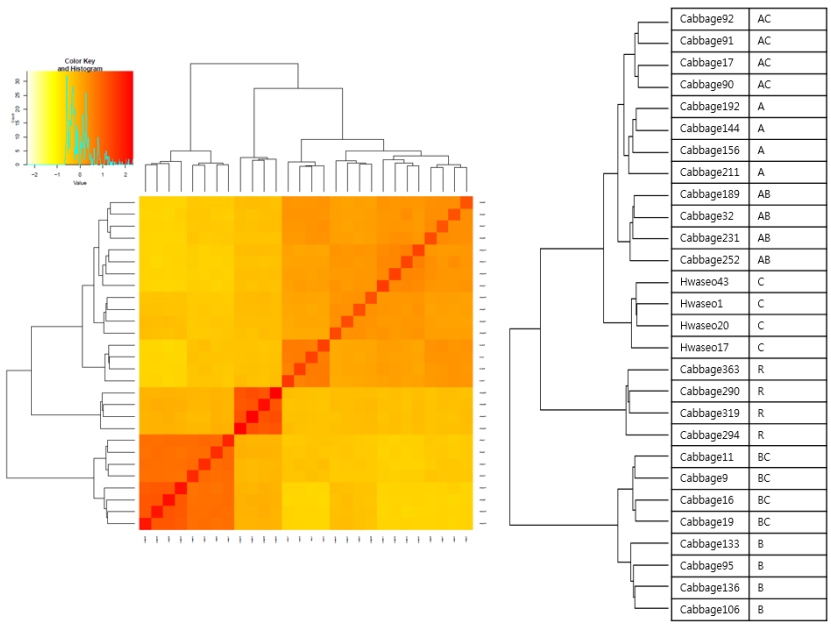

Figure 10. Heat map of a kinship matrix of 28 *Brassica* accessions based on 25 more SNPs genotyping in *A. thaliana*

It is assumed that this genetic relationship would correspond to the expected result. Even though the number of SNPs decreased with *A. thaliana* as reference genome, if the objective was to analyze selected SNP by minimizing some noise or bias, 3100 SNPs from diploid genome of 16 *Brassica* accessions and 3900 SNP from 28 *Brassica* accessions were not negligible figures. Consequently, this study implies that kinship analysis

method can lead to clear classification between individuals of plant species and to show its correlation genetically (which genome is more closed to another genome). In terms of relationship tree, this study can represent the evolutionary process in *Brassica* genus.

DISCUSSION

Advancement in data production and bioinformatics algorithms now make the variants analyses of complex polyploid genomes routine. This NGS based SNP data provides great opportunities to catalogue enormous natural variations within a population and interpret the information to understand the roles of adaptation and selection in diversity. Furthermore, the genetic variations are applied to identify trait associated molecular markers and predict breeding values for crop improvement. This, in turn, enables comparative genomic approaches to truly comprehend the effect of diversity on genome structure and how this impacts on the form and function of organisms, their growth and development, and response to environment, pests, and diseases.

SNP density varies between and within species, as well as different genomic regions. In rice, SNP density averages one in 147 bp (Subbaiyan et al. 2012), while soybean (Choi et al. 2007) and *Arabidopsis* (Atwell et al. 2010) average 1/438 and 1/500 bp, respectively. With the advent of Next-generation sequencing(NGS) technology, the discovery of large numbers of genome-wide SNPs is now highly achievable. Abundant markers can be discovered through amplicon sequencing, transcriptome sequencing, DNA-rich genome sequencing, and whole-genome sequencing (Henry 2012).

SNP prediction can be complicated by the error rate of NGS and by repetitive or highly homologous regions causing misassembly of short-read

lengths. However, the use of paired-end and large-insert NGS sequence reads in genome assembly and the strict quality control pipelines can help to minimize non-specific read mapping and false SNP predictions.

Given that the current public *Brassica* reference genome is limited to the *B. rapa, B. oleacea,* and *B. napus* genome, the primary limitation of SNP discovery from transcriptome sequencing or ESTbased sequencing is the restriction to coding regions and thus, this failed to detect diversity in non-coding regions. For genome-wide SNP discovery, low-coverage whole-genome resequencing is a simple, alternative approach for detecting polymorphisms in complex crop genomes by reducing the complexity of the genome.

In comparison of all four kinship analysis based on SNPs with multiple references (*A. thaliana, B. rapa, B. oleracea* and *B. napus*), each of the four accessions representing a genome type were grouped in a cluster regardless of reference genome. One distinct pattern observed was that the four accessions selected for each of *Raphanus* and *B. nigra* seemed to be more close to each other or less divergent from each other (big red square in Fig 10). Unlike phylogeny analysis mentioned above, *B. rapa* and *B. napus* as a reference, produced the same clustering patterns in kinship analysis and the results were similar to that of phylogenetic analysis based on whole genome SNP. Still, AB and BC tetraploid are not easy to place and the positions are dependent on the reference genome for analysis. AC tetraploid (*B. napus*) were close to A and C diploid, as expected. Figure 10 shows the heat map analysis of 28 accessions with four references.

Chapter III. Evolutionary Analysis of *Brassica* species

INTRODUCTION

Brassicaceae is one of the largest eudicot family which consists of more than 330 genera and 3800 species. It also comprises theeconomically most important crops that serve as a source of vegetables, oils, and fodders. The foundation for understanding the systematic relationship between the six major *Brassica* was laid by Korean-Japanese scientist Nagaharu U (U, 1935) and classically explained as U's triangle. He clearly proposed that the three tetraploid species *B. juncea* (AABB genome, 2n= 4x= 36), *B. napus* (AACC, 2n= 4x= 38), and *B. carinata* (BBCC, 2n= 4x= 34) are the derived allotetraploids of the diploid species *B. rapa* (AA, 2n= 2x= 20), *B. nigra* (BB, 2n= 2x= 16), and *B. oleracea* (CC, 2n= 2x= 18) caused by natural hybridization and chromosome doubling. Whole genome sequencing of the A, C, and AC genomes has increased our understanding of the *Brassica* genome evolution (Wang et al. 2011, Liu et al. 2014, Chalhoub et al. 2014, Parkin et al. 2014). It is suggested that *Brassica* genome was diverged from *A. thaliana*, the most extensively studied diploid model organism belonging to the same family of other *Brassica*, around 17 MYA (Yang et al. 2006b). Furthermore, the B genome was first diverged from *Brassica* lineage for about 9 MYA before A-C was diverged (around 4.5 MYA). It was expected that the AC-genome would diverge about 7000 years ago but there is no clear information about the other two tetraploid genomes, AB and BC (Chalhoub et al. 2014). A comparison with the close relatives, *Brassica* genome possess genome specific triplication, which is considered a crucial factor for evolution of morphotypes (Cheng et al. 2014). Due to its extensive variation in morphology and genome adaptation especially for A and C genomes, *Brassica* are considering to be a unique material for the polyploidy genome evolution studies. However, there has been no report about the ancestors of *Brassica* diploids. It is prerequisite to understand the

origin and evolution, and domestication history for further crop improvement (Allender and King, 2010).

Plant genome consists of three different evolutionary histories based on nuclear, mitochondrial, and chloroplast genomes. Among them, CpDNA and nuclear rDNA are the primary sources to understand the plant genome diversity and evolution, and compared to mitochondria, they possess highly informative phylogenetic signals due to its highly conserved nature (Kim et al. 2015b, Yang et al. 2015). CpDNA genomes are circular, relatively simple, uniparental inheritance mechanisms mostly via maternal inheritance (Birky, 1995, Reboud and Zeyl, 1994). CpDNA are highly conserved in gene structure and gene order with relatively diverse intergenic regions, due to its high conservation and low mutation rate CpDNA genomes are widely employed for understanding the genomic origin, genetic relationships, and barcoding marker development (Palmer et al. 1983, Moore et al. 2010, Li et al. 2015). In addition, CpDNA can provide population level genetic diversity, chloroplast inheritance, and phylogenetic relationships (Nikiforova et al. 2013, Zhang et al. 2016, Moore et al. 2010).

Nuclear ribosomal units are highly homozygous, tandemly-repeated transcriptional unit possessing important housekeeping role in nuclear assembly and nuclear function (Koo et al. 2011, Lim et al. 2005). Among two nrDNA blocks in plants (45S and 5S nrDNA) which are mostly localized on separate chromosomes, 45S nrDNA units are multi genic (18S, 5.8S, and 28S) possess high quality polymorphic information for phylogenetic and barcoding analysis (Waminal et al. 2015, Hasterok et al. 2001, Warwick et al. 2010).

NGS technology has remarkably increased our understanding of many genomes by decoding the genomic information (Varshney and May, 2012). Untill now, more than 100 whole genome draft assembly was generated for various plant species. The recent availability of the whole

genome sequences of A, C, and AC genome has explored the genome complexity, and uniqueness of the *Brassica* species for the study of polyploidy evolution as well as aid in crop improvement analysis (Chalhoub et al. 2014). NGS incursion rapidly increased the CpDNA genome studies. In addition, NGS advancement and arrival of NGS based tools help to obtain the complete CpDNA sequence more fast and accurate, so far >5000 complete CpDNA genome was available in the genbank including the 5 *Brassica* species (Seol et al. 2015, Hu et al. 2011). In addition, compare with CpDNA, very few reports on nrDNA of *Brassica* and the complete genome structure was not made available. Less than 100 nrDNA sequences were available at the genebank including two *Brassica* species (Waminal et al. 2015).

Furthermore, there have been various systematic studies on *Brassica* CpDNA for structure and diversity, but the studies were mostly limited and lacks comprehensive analysis of CpDNA genome of *Brassica* to understand the genetic relationship, origin, and diploid relationships which have yet to be resolved (Qiao et al. 2016, Sharma et al. 2014, Franzke et al. 2011). In this study, complete CpDNA and nrDNA sequences were generated for 28 *Brassica* and related species using low-coverage *de novo* assembly approach, and systematic phylogenic analysis was employed to address the maternal inheritance and genetic relationship and evolution of the U's triangle *Brassica* with its relative species. In addition, molecular divergence was estimated for the U's triangle *Brassica* based on the CpDNA and nrDNA. The sequencing and re-sequencing of different *Brassica* varieties has given researchers an unprecedented opportunity to identify genome wide variation. Tools in bioinformatics have been produced and were applied to interrogate and annotate this abundant data, and genome wide variation has been integrated with genetic maps and phenotypic information. The *Brassica* genomes, when combined with genome diversity information, provide an

insight into the evolution of these important crop plants and their wild relatives. Holistically, information on the whole genome scale and variants analysis gave insight into genome evolution and helps to develop species-specific barcoding markers.

MATERIALS AND METHODS

Plant materials and DNA sequencing

This study investigated the genetic diversity and evolution of *Brassica* species in U's triangle. Hence, researcher chooses seven different genotype groups among the *Brassica* species in order to conduct the comparative analysis between these lines (Table 3). Four germplasm from each each genome of A, B, C, R, AB, AC, and BC seeds were obtained from RDA-Genebank Information Center (http://www.genebank.go.kr/), Suwon, South Korea. All were grown in a farm at the Rural Development Administration (RDA) during spring season of 2014. Genomic DNA was extracted from approximately 5g samples of young leaves from all 28 genotypes, following the modified cetyltrimethylammonium bromide (CTAB) protocol (Allen et al. 2006). Prior to library preparation, the quality and quantity of the DNA were examined using both PicoGreen assay and NanoDrop ND-1000 (NanoDrop Technologies, Inc., USA) (Table 4). The multiplex identified adapters were used to separate the different accessions from the bulked raw reads.

Assembly of CpDNA and 45S rDNA cistron units

Complete CpDNA and nrDNA cistron units were simultaneously assembled for all the 28 accessions using dnalcw method (Table 10). Briefly,

raw reads of each accessions were processed under the clc-quality control tool to remove the adopter, low-quality sequences (with the phred score <20). The remaining high-quality reads were mapped onto to reference CpDNA genome (*A. thaliana*_NC00093) using clc_reference assembler to measure the mean coverage of CpDNA. In order to have more perfect assembly, sequencing reads containing about 100x coverage were used for the CpDNA assembly by clc-assembly cell package. Correction of errors such as false SNP, false gap, tandem repeat copy number, homo polymer copy number were carried out according to the dnalcw method (Kim et al. 2015b). Similarly, nrDNA units were assembled following the same approach for all the accessions. Due to the high variation into the intergenic sequences, even up to six kinds in *B. oleracea*, the unique 45S cistron units were only assembled. Moreover, the study was able to identify both parental types in a allotertaploid genome, *B. napus* (AC) containing parental or sub genomes *B. rapa* (A), and *B. oleracea* (C) represented as AC-A and AC-C, respectively.

Annotation of CpDNA and nrDNA

CpDNA from 28 accessions were annotated for protein-coding genes, transfer RNA (tRNA), and ribosomal RNA (rRNA) using Dual Organellar GenoMe Annotator(DOGMA) (Wyman et al. 2004). The accuracy of the start and stop codon and intron-exon boundaries were manually annotated based on the previously annotated information from its close relative (*A. thaliana*) which was used as a reference genome. tRNAscan-SE v1.2.1 was used further to validate the complete structure of the tRNA genes (Schattner et al. 2005). The systematic circular view of CpDNA was created using OGDRAW and in-house custom perl script (Lohse et al. 2007). Comparative syntenic map was using circos following the BlastZ annotation.

CpDNA based browser was developed for the systematic analysis of CpDNA genomes of 28 accessions and can be accessed at http://nabic.rda.go.kr/. CpBrowser also contains the sequence and gene annotation information for all the 28 accessions. Similarly, nrDNA genes (18S, 5.8S, and 26S) were annotated based on Blast analysis with reported reference units. The mVISTA tool was used to visualize the comparative syntenic relationship with other accessions. Complete sequences of CpDNA with annotation were stored at NABIC (Seol et al. 2016).

Structural Variants and PCR analysis

Inter- and intra- species structural variants such as SNPs, indels, and copy number variations were analyzed for 28 accessions. Putative single nucleotide variant (SNV) and indels were initially analyzed using Molecular Evolutionary Genetics Analysis version 7.0 (MEGA7). Tanden repeat (TR) finder was used to identify the tandem repeats. In order to detect highly reliable variants, all the predicted variation was manually curated for both CpDNA and nrDNA. Highly informative regions were validated by PCR analysis. In order to validate the polymorphic regions of CpDNA and nrDNA, specific primers were developed for structural high quality variants such as SNP and indels (Table 12, 16). DNA templates from 28 accessions were used for target fragment analysis. Each PCR reactions contains, 2 ng template DNA, 10 pM primers, 0.5 uM dNTPs, 2 units of Taq polymerase (TAKARA, Japan) with the final volume made up to 20ul with sterile distilled water. The PCR cycle consist of 10 min at 95℃ for pre-incubation, followed by 36 cycles of 30 secs at 94℃for incubation, 30 secs at 55-62℃ for annealing, 30 secs at 72℃ for initial extension, with 72℃ for 5 min for final extension. The amplicon was then checked under 2% agarose gel to access the estimated product size.

Phylogenetic and divergence time analysis

Complete CpDNA and 45S cistron units were separately employed for phylogenetic and divergence time estimation analysis. CpDNA sequences of 28 accessions with previously reported *Brassica* CpDNA were aligned by Multiple Alignment using Fast Fourier Transform (MAFFT) and phylogenetic tree was constructed using MEGA7 using neighbor-joining iterative model with 1000 bootstrap replications for more reliable tree. *A. thaliana* nrDNA sequence was used as an outgroup for the phylogenetic analysis of 40 nrDNA units based on 28 accessions (Kumar et al. 2016). The reference CpDNA sequence with its annotation of *A. thaliana, B. rapa, B. oleracea, B. nigra, B. juncea, B. carinata,* and *Raphanus sativus* were obtained from the genebank.

CpDNA and nrDNA sequences from 28 accessions were used for the tree topology and divergence time estimation using Bayesian method as implemented in Bayesian Evolutionary Analysis Sampling Trees (BEAST) program. BEAST program was widely used as a method that estimates the uncertainty of the divergence dates and branch length, also dating with known speciation dates, accommodate rate of among branches and includes the assumption of auto correlation.

BEAST program was used to compute the phylogeny tree and divergence time for major *Brassica* lineages using CpDNA and nrDNA sequences (Drummond et al. 2012). GTR+I+G substitution model was used to construct the tree topology and divergence time. It was performed for 10,000,000 generations of Markov-Chain-Monte-Carlo (MCMC) and sampled every 1000^{th} generations, with uncorrelated lognormal relaxed clock model, with Yule tree prior, and default prior for generated random starting tree. Tracer (v 1.6) was used to obtain BEAST run after discarding 10% generations as burn–in and the remaining BEAST runs were used for

the posterior possibilities. Tree annotator was used to estimate the divergence time. *A. thaliana* was constrained to be the outgroup and the age of divergence time between the *A. thaliana* and the *Brassica* lineage was constrained by a normal distribution with a mean of 17 MY and standard deviations of 2 MY (Yang et al. 2006a).

RESULTS

CpDNA and nrDNA sequence from *Brassica* and radish genotypes

Low coverage WGS (2-4x) was used to obtain the complete CpDNA and nrDNA of the major *Brassica* species and *R. sativus* by dnalcw approach. Annotation of CpDNA genomes has revealed conserved quadripartite structure with gene number and gene order among the 28 accessions (Figure 11, 16). Conservation of CpDNA genome was relatively high about 99-100% between accessions but very subtle variations were observed between the species (98.1-99.5). Genome size showed 607 bp variation among the 28 accessions with the ranges of 153,037 (A4) – 153,642 bp (B4), but in terms of copy numbers drastic variations was observed 453 (AB2)-1279 (BC1). Similarity analysis has revealed highly collinearity between the accessions (Figure 12). Likewise, compare with CpDNA, nrDNA showed highly conserved gene structure and size with little variations among the accessions. Overall, high level of synteny was observed for both CpDNA and nrDNA, especially nrDNA produced high syntenic relationship with *Brassica* and *Radish* genotypes (Figure 16).

Genetic diversity and variant analysis

Though the CpDNA and nrDNA are highly conserved because of its relatively simple structure and conserved gene structure and order, we observed considerable variation in both CpDNA and nrDNA genomes. CpDNA shows two diverse group, A and C genome, and derived allotetraploid (AB and AC) showed high similarity than compared with B, R genome and its allotetraploids (BC). Overall analysis of variable regions among and within the species has explored about 0.2% of diversity (Table 11). Extensive manual curation of CpDNA and nrDNA genomes reveals different kinds of non-redundant sequence variations (SV) such as SNV, indels and copy number variations. Among them SNV are the predominant one in both genome as 450 and 12 are present in CpDNA and nrDNA. Analysis with corresponding reference genome from genebank reveals more variations. Hence, the total number of SV are slightly higher than the observed for CpDNA genome. The average SNV density was 15/Kb and 3/Kb based on CpDNA and nrDNA, respectively. The distribution analysis of those variants reveals high proportion in intergenic regions than in genic regions. C-genome and R-genome showed low and high intra-species diversity based on CpDNA genome analysis (Figure 13). Likewise, ITS regions showed high variation then in genes of nrDNA in all the accessions and intra-species variations are rarely present compare with inter-species (Table 12) (Figure 16). Similar to CpDNA, more variations were observed between B and A-C lineage. Those most of the variations are intergenic, genic variation also observed for CpDNA genomes (Table 15). Especially, ycf2, atpB, atpB-rbcL, matK, ndhF, rbcL, rpl16, rps4-trnS, rps16, trnH-psbA, trnL-F, and trnS-G showed potential hotspot regions for barcoding marker development.

Comprehensive analysis of variable regions preset in the nrDNA allows us to develop barcoding markers for species identification. Using the single PCR reaction we can effectively discriminate all the four genomes (A, B, C, and R-genome) which will be highly a valuable information for molecular breeding and species identification (Table 14). In addition, validation of hotspot regions of CpDNA genome can able to differentiate each species separately (Figure 20).

Table 10. Summary statistics for assembly of CpDNA and 45S nrDNA sequences from 28 *Brassica* and related species

No	Sample ID.	Species	Genome size (Mb)	Total reads	Chloroplast genome			45S nrDNA			Genome [b]
					length (bp)	Mean Coverage (x)[a]	GC%	length (bp)	Mean Coverage (x) [a]	GC%	
1	144	*B. rapa*	529	1,556,496,282	153,483	378	36.36	5,818	3,216	53.08	A1
2	156	*B. rapa ssp. chinensis*	529	1,214,333,845	153,482	305	36.36	5,818	3,770	53.03	A2
3	192	*B. rapa ssp. pekinensis*	529	1,352,110,429	153,482	363	36.36	5,818	3,872	53.03	A3
4	211	*B. rapa ssp. rapa*	529	1,292,873,504	153,037	496	36.41	5,818	4,183	53.04	A4
5	95	*B. nigra*	632	1,532,334,602	153,633	378	36.39	5,831	1,819	53.32	B1
6	106	*B. nigra*	632	1,632,178,324	153,641	221	36.39	5,831	1,667	53.37	B2
7	133	*B. nigra*	632	1,488,961,884	153,623	323	36.39	5,831	1,324	53.34	B3
8	136	*B. nigra*	632	1,630,595,079	153,642	244	36.39	5,831	1,571	53.34	B4
9	h1	*B. oleracea ssp. capitata*	630	1,489,266,993	153,364	278	36.36	5,811	2,873	53.25	C1
10	h17	*B. oleracea ssp. botrytis*	630	1,312,276,810	153,364	510	36.36	5,848	1,384	53.25	C2
11	h20	*B. oleracea ssp. gongylodens*	630	1,610,969,185	153,364	285	36.36	5,818	2,768	53.30	C3
12	h43	*B. oleracea ssp. Italica*	630	2,114,898,597	153,363	347	36.36	5,819	1,957	53.23	C4
13	290	*Raphanus rapanisttrum ssp. landra*	530	1,466,836,940	153,372	264	36.34	5,816	3,812	53.08	R1
14	294	*Raphanus sativus var. raphanistroides*	530	1,487,196,488	153,444	412	36.32	5,816	2,042	53.13	R2
15	319	*Raphanus sativus var. sativus*	530	1,439,570,637	153,376	393	36.34	5,819	4,174	53.12	R3
16	363	*Raphanus sativus var. sativus*	530	1,469,683,462	153,370	343	36.34	5,823	4,614	53.12	R4

17	32	*B. juncea var. integrifolia*	1068	1,468,795,515	153,483	779	36.36	5,818	2,412	53.06	AB1-A
								5,831	1,589	53.42	AB1-B
18	189	*B. juncea var. integrifolia*	1068	1,352,110,429	153,483	358	36.36	5,818	1,883	53.06	AB2-A
								5,831	690	53.40	AB2-B
19	231	*B. juncea*	1068	1,527,999,376	153,490	495	36.36	5,818	2,192	53.06	AB3-A
								5,831	1,041	53.40	AB3-B
20	252	*B. juncea var. integrifolia*	1068	1,549,432,339	153,483	338	36.36	5,818	3,449	53.01	AB4-A
								5,831	1,190	53.40	AB4-B
21	9	*B. carinata*	1284	2,156,412,910	153,636	762	36.35	5,818	4,223	53.04	BC1-B
								5,818	2,409	53.08	BC1-C
22	11	*B. carinata*	1284	1,456,622,516	153,636	919	36.35	5,818	5,865	53.08	BC2-B
								5,818	3,453	53.28	BC2-C
23	16	*B. carinata*	1284	1,709,666,748	153,641	913	36.35	5,818	2,813	53.13	BC3-B
								5,817	1,836	53.17	BC3-C
24	19	*B. carinata*	1284	1,511,118,254	153,636	540	36.35	5,818	4,791	53.06	BC4-B
								5,818	2,551	53.18	BC4-C
25	17	*B. napus*	1130	1,534,168,684	153,452	630	36.39	5,831	1,445	53.35	AC1-A
								5,818	689	53.37	AC1-C
26	90	*B. napus var. napus*	1130	1,400,509,568	153,429	890	36.39	5,831	1,169	53.37	AC2-A
								5,819	879	53.17	AC2-C
27	91	*B. napus var. napus*	1130	1,400,509,568	153,429	925	36.39	5,817	1,009	53.45	AC3-A
								5,832	865	53.15	AC3-C
28	92	*B. napus var. napus*	1130	1,579,045,100	153,453	366	36.39	5,831	982	53.34	AC4-A
								5,818	741	53.27	AC4-C

[a] Copy numbers of CpDNA and nrDNA was estimated based on average read depth mapping and converted into their corresponding haploid genome size

[b] rDNA from tetraploid named as - A, - B, - C based on the parental genome source or sub-genome type

Table 11. Similarity and divergence plot based on 28 CpDNA sequences. The top triangle displays the similarity index to the maximum of 1. Yellow to green represents the low to high similarity index. The bottom triangle displays the number of nucleotide variable regions, while the red to blue boxes represents the low to high variations.

similarity/ variation	A1	A2	A3	A4	B1	B2	B3	B4	C1	C2	C3	C4	R1	R2	R3	R4	AB1	AB2	AB3	AB4	AC1	AC2	AC3	AC4	BC1	BC2	BC3	BC4
A1	ID	0.999	0.999	0.994	0.968	0.968	0.969	0.968	0.994	0.994	0.994	0.994	0.981	0.981	0.981	0.981	0.999	0.999	0.999	0.999	0.994	0.994	0.994	0.994	0.968	0.968	0.968	0.968
A2	16	ID	0.999	0.994	0.968	0.968	0.969	0.968	0.994	0.994	0.994	0.994	0.981	0.981	0.981	0.981	0.999	0.999	0.999	0.999	0.994	0.994	0.994	0.994	0.968	0.968	0.968	0.968
A3	8	12	ID	0.994	0.968	0.968	0.969	0.968	0.994	0.994	0.994	0.994	0.981	0.981	0.981	0.981	0.999	0.999	0.999	0.999	0.994	0.994	0.994	0.994	0.968	0.968	0.968	0.968
A4	775	775	775	ID	0.965	0.965	0.965	0.965	0.99	0.99	0.99	0.99	0.979	0.978	0.979	0.979	0.994	0.994	0.994	0.994	0.99	0.99	0.99	0.99	0.965	0.965	0.965	0.965
B1	4,812	4,812	4,811	5,348	ID	0.999	0.998	0.999	0.968	0.968	0.968	0.968	0.969	0.969	0.969	0.969	0.968	0.968	0.968	0.968	0.968	0.968	0.968	0.968	0.999	0.999	0.999	0.999
B2	4,811	4,812	4,808	5,350	47	ID	0.998	0.999	0.968	0.968	0.968	0.968	0.969	0.969	0.969	0.969	0.968	0.968	0.968	0.968	0.968	0.968	0.968	0.968	0.999	0.999	0.999	0.999
B3	4,774	4,777	4,773	5,323	205	198	ID	0.998	0.968	0.968	0.968	0.968	0.97	0.969	0.969	0.969	0.969	0.969	0.969	0.969	0.969	0.968	0.968	0.969	0.998	0.998	0.998	0.998
B4	4,813	4,814	4,810	5,352	51	12	200	ID	0.968	0.968	0.968	0.968	0.969	0.969	0.969	0.969	0.968	0.968	0.968	0.968	0.968	0.968	0.968	0.968	0.999	0.999	0.999	0.999
C1	915	914	912	1,518	4,914	4,918	4,892	4,915	ID	0.999	0.999	0.999	0.98	0.98	0.98	0.98	0.994	0.994	0.993	0.994	0.993	0.993	0.993	0.993	0.968	0.968	0.968	0.968
C2	908	909	907	1,512	4,907	4,911	4,883	4,914	10	ID	0.999	0.999	0.98	0.98	0.98	0.98	0.994	0.994	0.994	0.994	0.993	0.993	0.993	0.993	0.968	0.968	0.968	0.968
C3	910	909	907	1,513	4,909	4,913	4,887	4,910	5	9	ID	0.999	0.98	0.98	0.98	0.98	0.994	0.994	0.994	0.994	0.993	0.993	0.993	0.993	0.968	0.968	0.968	0.968
C4	908	909	907	1,512	4,907	4,911	4,883	4,914	11	1	10	ID	0.98	0.98	0.98	0.98	0.994	0.994	0.994	0.994	0.993	0.993	0.993	0.993	0.968	0.968	0.968	0.968
R1	2,803	2,811	2,806	3,156	4,640	4,642	4,621	4,643	2,944	2,937	2,939	2,936	ID	0.997	0.999	0.999	0.981	0.981	0.981	0.981	0.981	0.981	0.981	0.981	0.97	0.97	0.969	0.97
R2	2,906	2,914	2,909	3,253	4,726	4,728	4,697	4,722	3,027	3,026	3,022	3,025	437	ID	0.997	0.997	0.981	0.981	0.981	0.981	0.98	0.98	0.98	0.98	0.969	0.969	0.969	0.969
R3	2,833	2,839	2,834	3,174	4,675	4,677	4,646	4,679	2,973	2,966	2,968	2,965	99	432	ID	0.999	0.981	0.981	0.981	0.981	0.981	0.981	0.981	0.981	0.969	0.969	0.969	0.969
R4	2,827	2,833	2,828	3,168	4,669	4,671	4,640	4,673	2,967	2,960	2,962	2,959	93	426	8	ID	0.981	0.981	0.981	0.981	0.981	0.981	0.981	0.981	0.969	0.969	0.969	0.969
AB1	17	23	15	770	4,806	4,801	4,764	4,803	917	910	912	910	2,800	2,903	2,828	2,822	ID	0.999	0.999	0.999	0.994	0.994	0.994	0.994	0.968	0.968	0.968	0.968
AB2	18	22	14	771	4,805	4,800	4,765	4,802	916	911	911	911	2,800	2,903	2,828	2,822	3	ID	0.999	0.999	0.994	0.994	0.994	0.994	0.968	0.968	0.968	0.968
AB3	26	30	22	779	4,811	4,806	4,771	4,808	924	919	919	919	2,808	2,911	2,836	2,830	13	12	ID	0.999	0.994	0.994	0.994	0.994	0.968	0.968	0.968	0.968
AB4	18	22	14	771	4,805	4,800	4,765	4,802	916	911	911	911	2,800	2,903	2,828	2,822	3	2	12	ID	0.994	0.994	0.994	0.994	0.968	0.968	0.968	0.968
AC1	794	798	794	1,400	4,819	4,815	4,783	4,808	947	948	946	949	2,839	2,934	2,878	2,872	791	792	799	792	ID	0.999	0.999	0.999	0.968	0.968	0.968	0.968
AC2	816	820	816	1,422	4,839	4,835	4,803	4,828	973	974	970	975	2,861	2,956	2,900	2,894	813	814	821	814	26	ID	1	0.999	0.968	0.968	0.968	0.968
AC3	816	820	816	1,422	4,839	4,835	4,803	4,828	973	974	970	975	2,861	2,956	2,900	2,894	813	814	821	814	26	-	ID	0.999	0.968	0.968	0.968	0.968
AC4	789	793	789	1,395	4,814	4,810	4,778	4,811	953	948	950	949	2,834	2,937	2,873	2,867	786	787	794	787	7	29	29	ID	0.968	0.968	0.968	0.968
BC1	4,807	4,808	4,804	5,346	53	6	194	18	4,914	4,907	4,909	4,907	4,638	4,724	4,673	4,667	4,797	4,796	4,802	4,796	4,811	4,831	4,831	4,806	ID	1	0.999	1
BC2	4,807	4,808	4,804	5,346	53	6	194	18	4,914	4,907	4,909	4,907	4,638	4,724	4,673	4,667	4,797	4,796	4,802	4,796	4,811	4,831	4,831	4,806	-	ID	0.999	1
BC3	4,812	4,813	4,809	5,351	48	1	197	11	4,919	4,912	4,914	4,912	4,642	4,729	4,678	4,672	4,802	4,801	4,807	4,801	4,815	4,835	4,835	4,810	7	7	ID	0.999
BC4	4,807	4,808	4,804	5,346	53	6	194	18	4,914	4,907	4,909	4,907	4,638	4,724	4,673	4,667	4,797	4,796	4,802	4,796	4,811	4,831	4,831	4,806	-	-	7	ID

Table 12. Summary of nucleotide variations based on the 40 nrDNA units from 28 accessions

	group-specific	C		B	R	A=C/B=R	A	A/B/C/R	A	B	A	B	B	R	R	B	A=C/B=R				R	R	R
genome	Position	498	671	691	699	716	1859	1873	1883	1972	2006	2335	2350	2877	2924	2948	2953	3167	3168	3188	3189	4420	4430	4473
AA	A1_1	T	T	T	C	C	A	-------	T	A	A	A	T	C	G	T	A	C	T	A	G	A	G	T
	A2_1	T	T	T	C	C	A	-------	T	A	A	A	T	C	G	T	A	C	T	A	G	A	G	T
	A3_1	T	T	T	C	C	A	-------	T	A	A	A	T	C	G	T	A	C	T	A	G	A	G	T
	A4_1	T	T	T	C	C	A	-------	T	A	A	A	T	C	G	T	A	C	T	A	G	A	G	T
BB	B1_1	T	C	C	C	T	G	GCCGATT	C	T	C	G	C	C	G	C	G	T	A	T	A	A	G	T
	B2_1	T	C	C	C	T	G	GCCGATT	C	T	C	G	C	C	G	C	G	T	A	T	A	A	G	T
	B3_1	T	C	C	C	T	G	GCCGATT	C	T	C	G	C	C	G	C	G	T	A	T	A	A	G	T
	B4_1	T	C	C	C	T	G	GCCGATT	C	T	C	G	C	C	G	C	G	T	A	T	A	A	G	T
CC	C1_1	C	T	T	C	C	G	GCTGATT	C	A	C	A	T	C	G	T	A	C	T	A	G	A	G	T
	C1_2	C	C	T	C	C	G	GCTGATT	C	A	C	A	T	C	G	T	A	C	T	A	G	A	G	T
	C2_1	C	T	T	C	C	G	GCTGATT	C	A	C	A	T	C	G	T	A	C	T	A	G	A	G	T
	C2_2	C	T	T	C	C	G	GCTGATT	C	A	C	A	T	C	G	T	A	C	T	A	G	A	G	T
	C3_1	C	T	T	C	C	G	GCTGATT	C	A	C	A	T	C	G	T	A	C	T	A	G	A	G	T
	C3_2	C	C	T	C	C	G	GCTGATT	C	A	C	A	T	C	G	T	A	C	T	A	G	A	G	T
	C4_1	C	T	T	C	C	G	GCTGATT	C	A	C	A	T	C	G	T	A	C	T	A	G	A	G	T
	C4_2	C	C	T	C	C	G	GCTGATT	C	A	C	A	T	C	G	T	A	C	T	A	G	A	G	T
RR	R1_1	T	C	T	T	T	G	CCGGAAT	C	A	C	A	T	T	A	T	G	T	A	T	A	G	A	C
	R2_1	T	C	T	T	T	G	CCGGAAT	C	A	C	A	T	T	A	T	G	T	A	T	A	G	A	C
	R3_1	T	C	T	T	T	G	CCGGAAT	C	A	C	A	T	T	A	T	G	T	A	T	A	G	A	C
	R4_1	T	C	T	T	T	G	CCGGAAT	C	A	C	A	T	T	A	T	G	T	A	T	A	G	A	C

AABB	AB1_1	T	T	T	C	C	A	-------	T	A	A	A	T	C	G	T	A	C	T	A	G	A	G	T
	AB1_2	T	C	C	C	T	G	GCCGATT	C	T	C	G	C	C	G	C	G	T	A	T	A	A	G	T
	AB2_1	T	T	T	C	C	A	-------	T	A	A	A	T	C	G	T	A	C	T	A	G	A	G	T
	AB2_2	T	C	C	C	T	G	GCCGATT	C	T	C	G	C	C	G	C	G	T	A	T	A	A	G	T
	AB3_1	T	T	T	C	C	A	-------	T	A	A	A	T	C	G	T	A	C	T	A	G	A	G	T
	AB3_2	T	C	C	C	T	G	GCCGATT	C	T	C	G	C	C	G	C	G	T	A	T	A	A	G	T
	AB4_1	T	T	T	C	C	A	-------	T	A	A	A	T	C	G	T	A	C	T	A	G	A	G	T
	AB4_2	T	C	C	C	T	G	GCCGATT	C	T	C	G	C	C	G	C	G	T	A	T	A	A	G	T
AACC	AC1_1	T	T	T	C	C	A	-------	T	A	A	A	T	C	G	T	A	C	T	A	G	A	G	T
	AC1_2	C	T	T	C	C	G	GCTGATT	C	A	C	A	T	C	G	T	A	C	T	A	G	A	G	T
	AC2_1	T	T	T	C	C	A	-------	T	A	A	A	T	C	G	T	A	C	T	A	G	A	G	T
	AC2_2	C	C	T	C	C	G	GCTGATT	C	A	C	A	T	C	G	T	A	C	T	A	G	A	G	T
	AC3_1	T	C	T	C	C	A	-------	T	A	A	A	T	C	G	T	A	C	T	A	G	A	G	T
	AC3_2	C	T	T	C	C	G	GCTGATT	C	A	C	A	T	C	G	T	A	C	T	A	G	A	G	T
	AC4_1	T	T	T	C	C	A	-------	T	A	A	A	T	C	G	T	A	C	T	A	G	A	G	T
	AC4_2	C	T	T	C	C	G	GCTGATT	C	A	C	A	T	C	G	T	A	C	T	A	G	A	G	T
BBCC	BC1_1	T	C	C	C	T	G	GCCGATT	C	T	C	G	C	C	G	C	G	T	A	T	A	A	G	T
	BC1_2	C	T	T	C	C	G	GCTGATT	C	A	C	A	T	C	G	T	A	C	T	A	G	A	G	T
	BC2_1	T	T	T	C	C	G	GCCGATT	C	T	C	G	C	C	G	C	G	T	A	T	A	A	G	T
	BC2_2	C	C	C	C	T	G	GCTGATT	C	A	C	A	T	C	G	C	G	C	T	A	G	A	G	T
	BC3_1	T	C	C	C	T	G	GCCGATT	C	T	C	G	C	C	G	C	G	T	A	T	A	A	G	T
	BC3_2	C	T	T	C	C	G	GCTGATT	C	A	C	A	T	C	G	T	A	C	T	A	G	A	G	T
	BC4_1	T	C	C	C	T	G	GCCGATT	C	T	C	G	C	C	G	C	G	T	A	T	A	A	G	T
	BC4_2	C	T	T	C	C	G	GCTGATT	C	A	C	A	T	C	G	T	A	C	T	A	G	A	G	T

Table 13. nrDNA copy number variations among the sub-genomes in *Brassica* tetraploids

species	AB1		AB2		AB3		AB4	
nrDNA type	1-A	2-B	1-A	2-B	1-A	2-B	1-A	2-B
depth (X)	3,317	2,186	2,384	873	3,136	1,489	5,004	1,727
ratio	1.5	1	2.7	1	2.1	1	2.9	1

species	AC1		AC2		AC3		AC4	
nrDNA type	1-A	2-C	1-A	2-C	1-A	2-C	1-A	2-C
depth (X)	7,092	4,045	6,654	3,917	3,745	2,445	5,638	3,002
ratio	1.8	1	1.7	1	1.5	1	1.9	1

species	BC1		BC2		BC3		BC4	
nrDNA type	1-B	2-C	1-B	2-C	1-B	2-C	1-B	2-C
depth (X)	1,962	936	1,449	1,090	1,250	1,072	1,372	1,036
ratio	2.1	1	1.3	1	1.2	1	1.3	1

Table 14. nrDNA sub-genome dominance

Primer no.	Kinds	Specificity	Locus	Locus	direction	Sequence (5'-3')	Primer length	Product size (sample no. - bp) (pcr application)
primer 01	1(A,C) 2(AB,AC,B,BC), 3(R)	group-specific	trnH-GUG	psbA	01 F :	CCATCGAAGAGAAGCAAATGA	21	1(239), 2(238), 3(231)
	R	C294			01 R :	CCTCTCGGGGACTTGCTTA	19	(C294: 230) HRM (high resolution melting)
primer 02	A	C211	trnS-GCU	trnR-UCU	02 F :	TCCACTCAGCCATCTCTCCT	20	C211: 389
					02 R :	TTCTCTTTCGAGCCTTTTCTTT	22	C144, 156, 192: 372
primer 03	A	C156	ycf3		03 F :	CTAAATTTCCAGGAATTAGTCAC	23	C156: 206
					03 R :	TACGAATAGGAGGCACAGGG	20	C211, 144, 192: 188
primer 04	AB	C32	rbcL	accD	04 F :	CGGAGTTCCACCTGAAGAAG	20	C32: 219
					04 R :	GAGGTAAACATGTTAGTAACAGAC	24	C189, 252, 231: 207 HRM
primer 05	AB	C252	trnS-GCU	trnR-UCU	05 F :	CGCCTTTTCTATCTTCTAGA	20	C252: 262
					05 R :	CATCGTTAGCTTGGAAGGCT	20	C32, 189, 231: 247
primer 06	AB	C231	trnS-GCU	trnR-UCU	06 F :	AAACCCTTAGCCTTCCAAGC	20	C32, 189, 252: 266
					06 R :	TGCGTCCAATAGGATTTGAA	20	C231: 273
primer 07	1(AC), 2(A,AB,B,BC,C,R)	group-specific	trnT-GGU	psbD	07 F :	TAACTCAGTGGTAGAGTAACGCC	23	1(395), 2(344)
					07 R :	TTCGACGGGTTAGGTCCAC	19	
primer 08	AC	C92	trnH-GUG	psbA	08 F :	AAAAATGATTTGTTCCGTTTTATAG	25	C92: A x 4
					08 R :	CGTGCTAACCTTGGTATGGAA	21	C17, 90, 91: T x 4 HRM
primer 09	C	H20	psbK	psbI	09 F :	TGGATCATTTGATTTCCTCAGTT	23	H20: A
					09 R :	GTAATCCGGGACGTGAAGAA	20	H1, 17, 43: T HRM
primer 10	C	H1	trnV-UAC		10 F :	CCCTACCGAAATGGGGTACT	20	H1: A x 7
					10 R :	TGGATCATAAACACAAGGGCTA	22	H20, 17, 43: A x 8 HRM
primer 11	C	H43	petA	psbJ	11 F :	ATTGTGTCAGTCGGGAAAGC	20	H43: C x 10
					11 R :	GGCCCAACTCTTCCTTCTTT	20	H1, 20, 17: C x 11 HRM
primer 12	B	C133	rps16	trnQ-UUG	12 F :	CATGAATAGTCATAGTTCAGCCAGT	25	C133: 372
					12 R :	TTATTTCAACCGAAATTACAAAAA	24	C95, 136, 106: 381
primer 13	B	C95	psbM	trnD-GUC	13 F :	CCTTGGTGGGATTCGAACTA	20	C95: 305
					13 R :	AGGGGATCAAAATGGTTTCG	20	C133, 136, 106: 310 HRM
primer 14	1(A,AB) 2(AC) 3(B,BC) 4(C) 5(R)	group-specific	trnH-GUG	psbA	14 F :	TCCACTGCCTTAATCCACTTGG	22	1(381), 2(386), 3(312), 4(380), 5(365)
	A / B	C133			14 R :	CCGTGCTAACCTTGGTATGG	20	(C211: 380, C133: 317) HRM
primer 15	1(A,AB) 2(B,BC) 3(AC) 4(R) 5(C)	group-specific	trnK-UUU	rps16	15 F :	CCAGTCATGTGTGCGTCAGG	20	1(514), 2(516), 3(481), 5(488)
	R	C290, C249			15 R :	CTACTCTTTTCTTTCCTCCTC	20	C290: 559 C319,C363: 549 C249: 604
primer 16	1(A,AB,C) 2(B,BC) 3(AC) 4(R)	group-specific	rps16	trnQ-UUG	16 F :	TCCTTCAATTCAAGTCGCACG	21	1(539), 2(538), 3(543), 4(548)
	B / R	C294			16 R :	GGTTCGAATCCTTCCGTCCC	20	(C133: 529, C294: 530)
primer 17	1(A,AB) 2(BC) 3(AC) 4(R) 5(C)	group-specific	trnS-GCU	trnR-UCU	17 F :	TAGTCCACTCAGCCATCTCTC	21	1(478), 2(475), 3(495), 4(452), 5(456)
	A / B / R	C211			17 R :	CTGACCAGGCCAGGCTAT	18	(C211: 495, C106: 475 C136: 460 C95: 476 C133: 478, C:294: 459)
primer 18	1(A,AB,AC,R) 2(B,BC), 3(C)	group-specific	atpF		18 F :	ACGTAGGTTATCAATTCTGCATTA	24	1(502), 2(503), 3(508) HRM
					18 R :	TACTTGGGTCACTGGCCATC	20	
primer 19	1(A,AB,C) 2(B,BC) 3(AC) 4(R)	group-specific	atpH	atpI	19 F :	GATACCTTCGACAGCTTGAC	20	1(664), 2(722), 3(665), 4(665)
(수정)	B / R				19 R :	TTTACAAGCGGGATTCAAGC	20	(C95: 515, C290: 457 C319,363: 456 C294: 455)
primer 20	1(A,AB) 2(B) 3(BC) 4(AC,C,R)	group-specific	trnC-GCA	petN	20 F :	CCTGGCTCTCGAGGTTCTATT	21	1(387), 2(351), 3(346), 4(347)
	A / B / BC	C133, C16			20 R :	AGAATCGACAAGATGTAAACACAA	24	(C211: 367, C133: 346, C16: 351) HRM
primer 21	1(A,AB,C) 2(B,BC) 3(AC) 4(R)	group-specific	ycf3		21 F :	ACCTCATACGGCTCGACAAC	20	1(404), 2(361), 3(409), 4(411)
	R	C294			21 R :	AAGGTATTGAGCAGCGGTGT	20	(C294: 423)
primer 22	1(A,AB) 2(B,BC) 3(AC) 4(C) 5(R)	group-specific	trnF-GAA	ndhJ	22 F :	GCTCAGTTGGTAGAGCAGAGG	21	1(575), 2(500), 3(521), 4(518), 5(509)
	A / B / AC	C211, C17,C92			22 R :	TCCTAAAAGCCGAGCCAATA	20	(C211: 629, C95: 499, C17,92: 521 C90,91: 498)
primer 23	1(A,AB,AC) 2(B,BC) 3(C) 4(R)	group-specific	petA	psbJ	23 F :	TCTGTCGTTAGTGACCAATTGAA	23	1(256), 2(255), 3(300), R(250)
	R	C294			23 R :	TGCAATTAAGAATGACAAGATCG	23	(C294: 255) HRM

Phylogenetic analysis of U's triangle *Brassica* with its relatives

Complete CpDNA and nrDNA sequences from 28 accessions were used for the phylogenetic analysis with MEGA7 (Figure 14, 17). The phylogeny obtained with two different genome shows almost identical tree topology. However, variations were observed for AC genome and intra-species level. In both case *A. thaliana* was used as an outgroup for rooted phylogeny. The high bootstrap values on the node shows the reliability of the phylogeny recovered from both CpDNA and nrDNA (Figure 14, 17). The phylogeny based on CpDNA displayed five different clades with clear discrimination of four diploid genomes with an ambiguous clade based on

AC-genome (Figure 14). The other tetraploids (AB, BC) were following their parents, either the A or B genome. However, no ambiguous clade was observed when the phylogeny based on the nrDNA sequence (Figure 17). The four major clades were monophyletic with their parental diploid species and the tetraploid genome follows their corresponding progenitor genome, such as AB follows A and B, BC follows B and C, and AC follows A and B. Compare with other genome AC are expected to show high diversity within and among the species. It is important to note that, any reciprocal hybridization pattern in all the three tetraploids was never observed.

Phylogenetic analysis based on CpDNA and nrDNA showed conserved pattern genetic relationship among the seven species survived. In addition, phylogeny revealed the maternal/paternal inheritance of the three allotetraploid *Brassica* species. Though clades are monophyletic intra species divergence was observed for some CpDNA (A4, B3, R2) and nrDNA (AC2-A, AC2-C) as well. Overall, the genetic relation of *Brassica* species follows the general trend with the previously reported as B genome are sister group to R, A, and C genome and the derived allotetraploids followed their corresponding progenitor genomes.

Table 15. Chloroplast genic variations between the *B. rapa* accessions

locus	type	c211	c144	c156	c192	site
matK	SNP	T	A	A	A	3498
matK	2bp INDEL	--	AT	AT	AT	3756 - 3757
rps2	SNP	A	G	G	G	14730
rps2	SNP	C	A	A	A	14769
rpoC2	SNP	A	C	A	A	17440
rpoC2	SNP	A	G	G	G	17689
rpoB	SNP	G	T	T	T	24164
rpoB	SNP	A	A	G	A	25698
psaB	SNP	T	G	G	G	36927
psaA	SNP	C	T	T	T	40087
psaA	SNP	G	T	T	T	40800
ndhK	SNP	A	G	G	G	48520
accD	INDEL	6bp	-	-	-	56551 - 56556
ycf2	SNP	G	T	T	T	86916
ndhH	SNP	T	G	G	G	120808

Divergence time estimation in the Genus *Brassica*

Complete CpDNA and nrDNA sequence based independent estimation of speciation time Bayesian method implemented in BEAST program (Figure 15). Tree topology developed by the BEAST analysis similar with the phylogeny created using MEGA, demonstrate the high quality and reproducibility of our phylogeny. The molecular dating based on CpDNA and nrDNA showed almost similar divergence time and the results are shown in Figure 15. The divergence time estimated for all the 7 species based on CpDNA and nrDNA reveals clear and major speciation events for the four diploids then the tetraploids. For both CpDNA and nrDNA tree topology, *A. thaliana* was used as an outgroup and was assumed that the divergence time of *Brassica* lineage with *A. thaliana* was about 17 MYA. The BEAST analysis is a statistical model which used high precision of divergence time, and the estimate was compared with the previous estimation. Tree topology with inferred speciation dates clearly shown three major period of divergence in both analysis (CpDNA and nrDNA) for the *Brassica* genus. According to the results, the tree demonstrates that the divergence and speciation of B genome ocurred 11 MYA. After which, about 5.4 MYA, the R genome was independently diverged from the B genome. Later, about 4.5 MYA, A and C genome speciation ocurred from the B genome. The natural allopolyploidization producing AB, BC, and AC genome and estimation of divergence time for three allopolyploids are quite ambiguous. In general, all the allotetrploids showed high difference rate of divergence even with the species and are expected to be derived from its diploid ancestor during 0.001 to 0.03 MYA (Figure 19). The estimation of recent divergence time seems to be influenced by artificial hybridization, domestication, and random crossing during historical events.

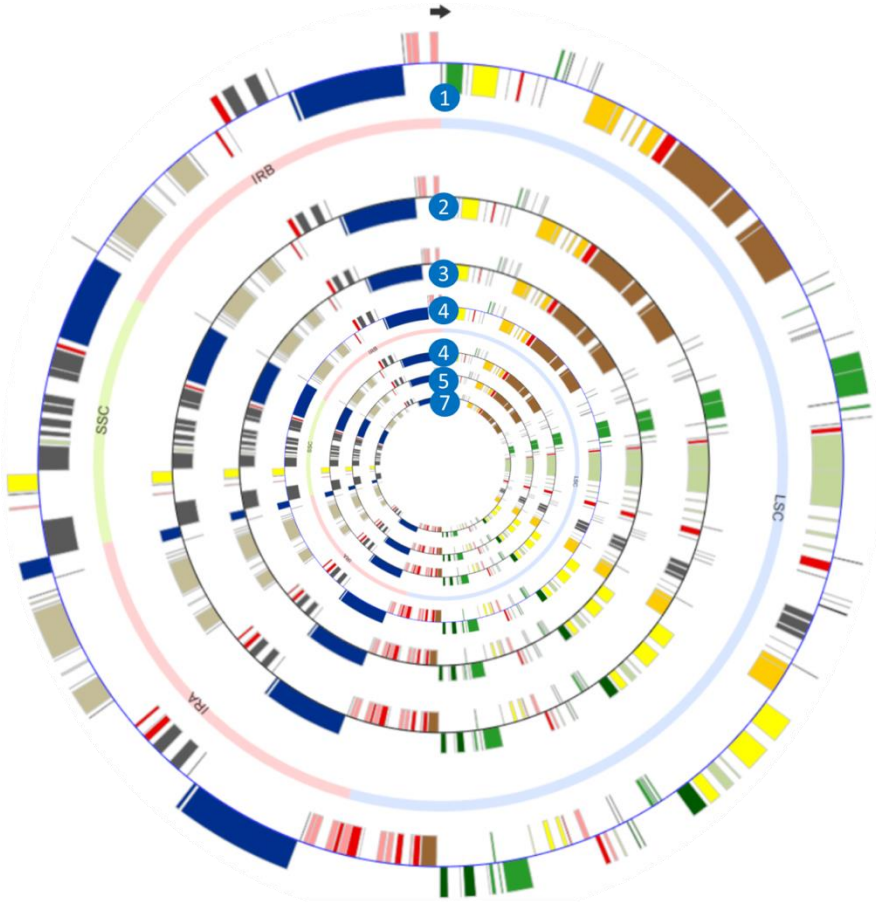

Figure 11. Circular map of the chloroplast genome from seven species belonging to the Brassicaceae family. *B. napus* (1), *B. carinata* (2), *B. Juncea* (3), *Raphanus sativus* (4), *B. oleraca* (5), *B. nigra* (6), *and B. rapa* (7). The genes are represented in different color bars. The positive and negative orientation of the genes are shown as outer and inner circle bars. The two inverted repeat region (IRA and IRB) flanked by large sub unit (LSC) and small sub unit (SSC) are represented between first and second circular map.

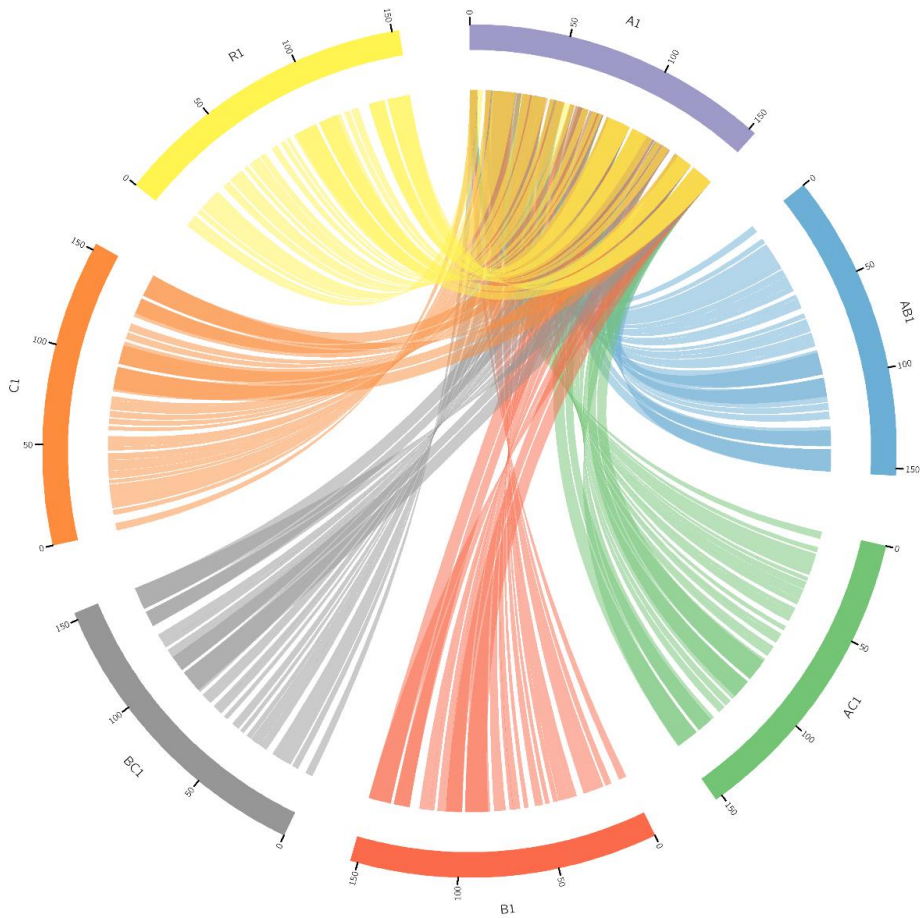

Figure 12. Synteny comparisons of chloroplast genomes from the genus *Brassica*. A circos based syntenic comparative map developed for *B. rapa* (A1) against *B. Juncea* (AB1), *B. napus* (AC1), *B. nigra* (B1), *B. carinata* (BC1), *B. oleraca* (C1), and *Raphanus sativus* (R1). Syntenic block with the minimum of 1 Kb length was used for the syntenic analysis.

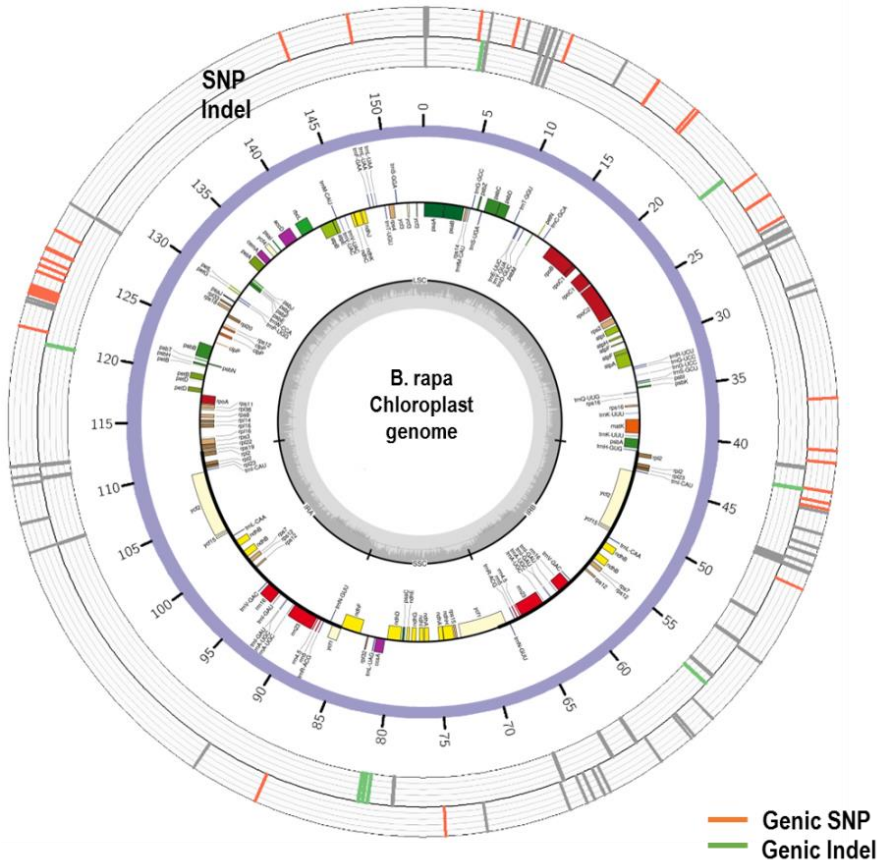

Figure 13. Distribution of intra-species variation of *B. rapa* chloroplast genome. The outer and inner circle represents SNP and indels, respectively. Color bars on the inner and outer circle represents genic variations. Innermost chloroplast circular map was developed using OGDRAW.

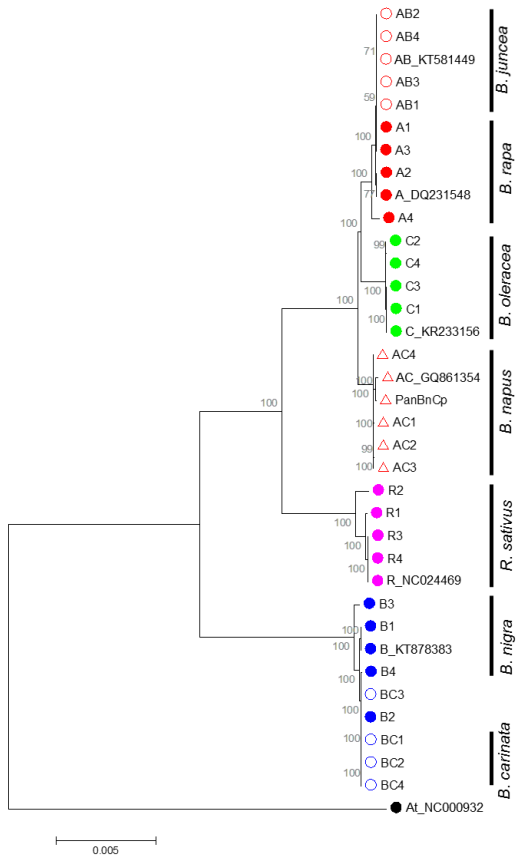

Figure 14. Phylogenetic relationships of the genus *Brassica*. The neighbor-joining tree inferred from complete CpDNA from 28 accessions. Tree wad developed using MEGA7 with 1000 bootstrap replications. The bootstrap values for clades are shown in corresponding branches of the tree. Filled and unfilled circles (legends) represents diploid and tetraploids accessions, respectively. Reference genomes (accessions with genbank number) are obtained from NCBI.

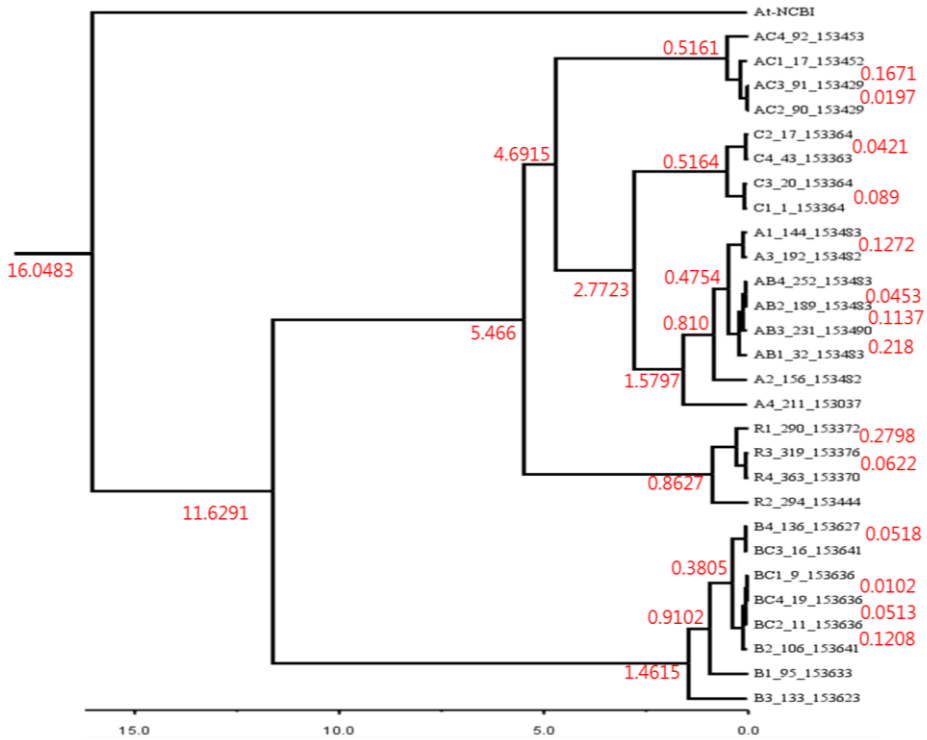

Figure 15. Chronogram of *Brassica* genus inferred from Bayesian analysis as implemented in BEAST program based on complete cp genome. Divergence time of species are on the right side node represented as MY. The tree and dating were done according to the protocol mentioned in the material and methods section.

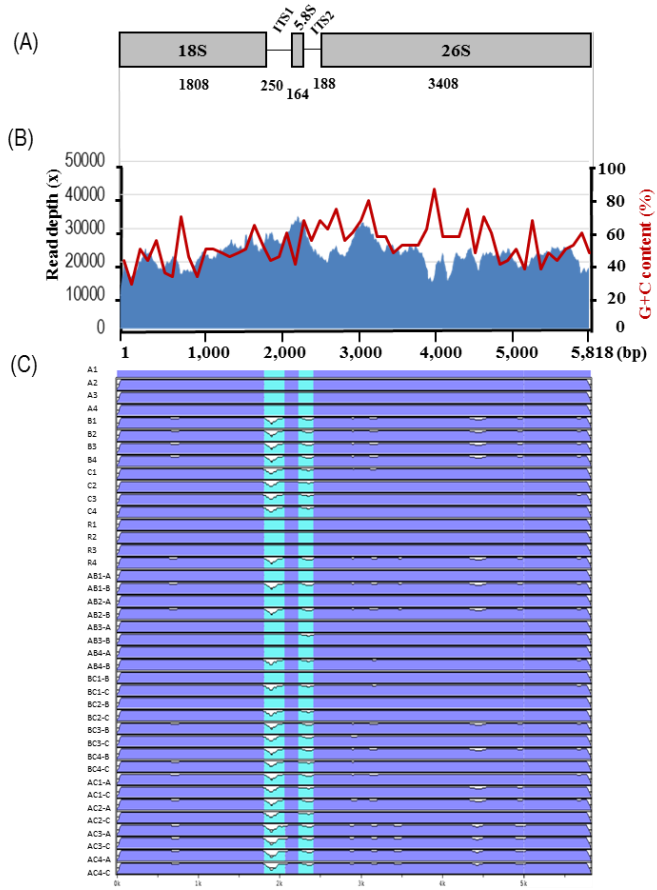

Figure 16. Structure and similarity analysis of 45s nrDNA cistron based on 28 accessions. (A) complete structure and gene annotation of 45S rDNA cistron unit of *B. rapa*. (B) coverage of 45 nrDNA based read mapping. Read line graph indicates the G+C proportion of the 45S nrDNA. (C) mVISTA based comparative analysis displayed the similarity and variable regions among the 28 accessions.

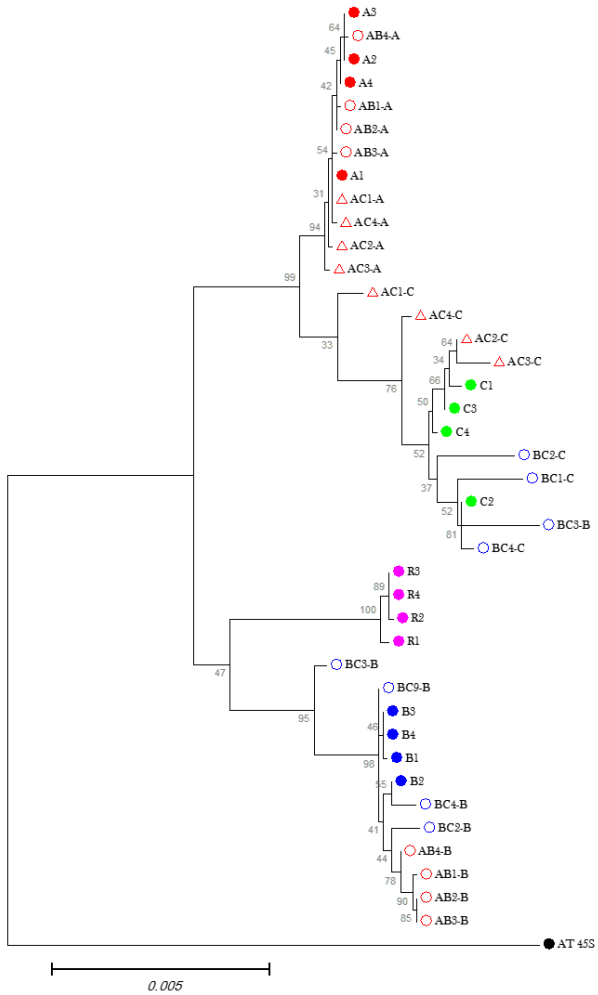

Figure 17. Phylogenetic analysis based on nrDNA. The neighbor-joining tree inferred from 45s rDNA cistron units from 28 accessions. Tree was developed using MEGA7 with 1000 bootstrap replications. The bootstrap values for clades are shown in corresponding branches of the tree. Filled and unfilled circles (legends) represents diploid and tetraploids accessions, respectively. *A. thaliana* was used as an outgroup.

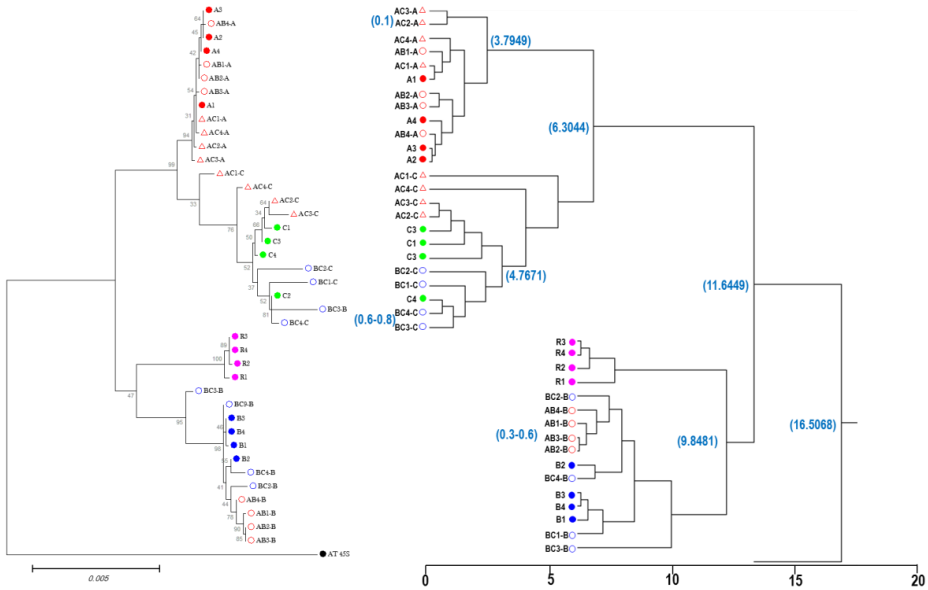

Figure 18. Phylogenetic relationships and molecular dating of the genus *Brassica* based on nrDNA. (left tree) The neighbor-joining tree inferred from complete nrDNA from 28 accessions. Tree was developed using MEGA7 with 1000 bootstrap replications. The bootstrap values for clades are shown in corresponding branches of the tree. Filled and unfilled circles (right tree). Divergence time of species are on the right side node represented as MY. The tree and dating were done according to the protocol mentioned in material and methods section. Blue bars indicate the error range for corresponding branching points.

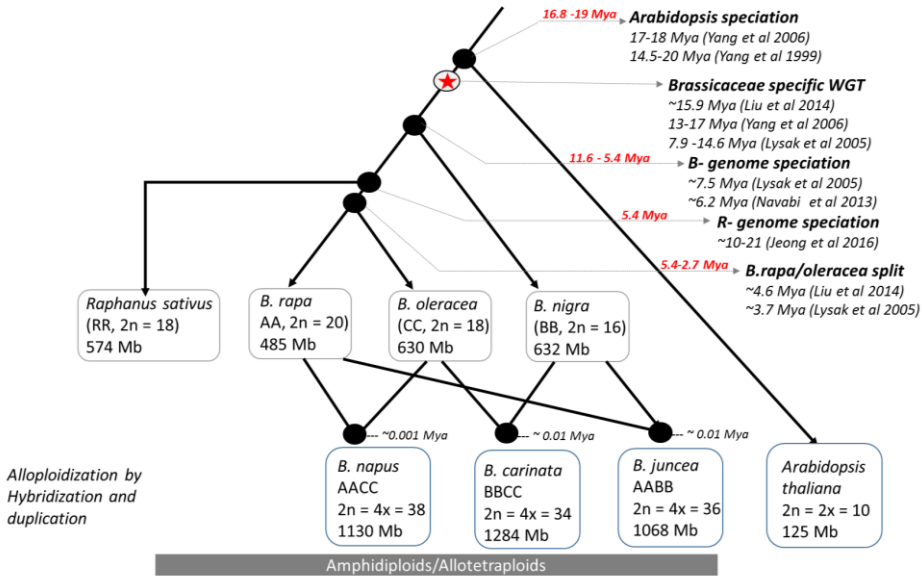

Figure 19. Evolution of *Brassica* species

Table 16. Chloroplast oligonucleotide primers used in the study

Primer no.	Kinds(vs)	Description	direction	Sequence (5'-3')
primer 01	1(A,C) 2(AB,AC,B,BC), 3(R)	group-specific	01 F :	CCATCGAAGAGAAGCAAATGA
			01 R :	CCTCTCGGGGACTTGCTTA
primer 07	1(AC), 2(A,AB,B,BC,C,R)	group-specific	07 F :	TAACTCAGTGGTAGAGTAACGCC
			07 R :	TTCGACGGGTTAGGTCCAC
primer 14	1(A,AB) 2(AC) 3(B,BC) 4(C) 5(R)	group-specific	14 F :	TCCACTGCCTTAATCCACTTGG
			14 R :	CCGTGCTAACCTTGGTATGG
primer 15	1(A,AB) 2(B,BC) 3(AC) 4(R) 5(C)	group-specific	15 F :	CCAGTCATGTGTGCGTCAGG
			15 R :	CTACTCTTTTCTTTCCTCCTC
primer 16	1(A,AB,C) 2(B,BC) 3(AC) 4(R)	group-specific	16 F :	TCCTTCAATTCAAGTCGCACG
			16 R :	GGTTCGAATCCTTCCGTCCC
primer 17	1(A,AB) 2(BC) 3(AC) 4(R) 5(C)	group-specific	17 F :	TAGTCCACTCAGCCATCTCTC
			17 R :	CTGACCAGGCCAGGCTAT
primer 18	1(A,AB,AC,R) 2(B,BC), 3(C)	group-specific	18 F :	ACGTAGGTTATCAATTCTGCATTA
			18 R :	TACTTGGGTCACTGGCCATC
primer 19	1(A,AB,C) 2(B,BC) 3(AC) 4(R)	group-specific	19 F :	GATACCTTCGACAGCTTGAC(수정)
			19 R :	TTTACAAGCGGGATTCAAGC
primer 20	1(A,AB) 2(B) 3(BC) 4(AC,C,R)	group-specific	20 F :	CCTGGCTCTCGAGGTTCTATT
			20 R :	AGAATCGACAAGATGTAAACACAA
primer 21	1(A,AB,C) 2(B,BC) 3(AC) 4(R)	group-specific	21 F :	ACCTCATACGGCTCGACAAC
			21 R :	AAGGTATTGAGCAGCGGTGT
primer 22	1(A,AB) 2(B,BC) 3(AC) 4(C) 5(R)	group-specific	22 F :	GCTCAGTTGGTAGAGCAGAGG
			22 R :	TCCTAAAAGCCGAGCCAATA
primer 23	1(A,AB,AC) 2(B,BC) 3(C) 4(R)	group-specific	23 F :	TCTGTCGTTAGTGACCAATTGAA
			23 R :	TGCAATTAAGAATGACAAGATCG

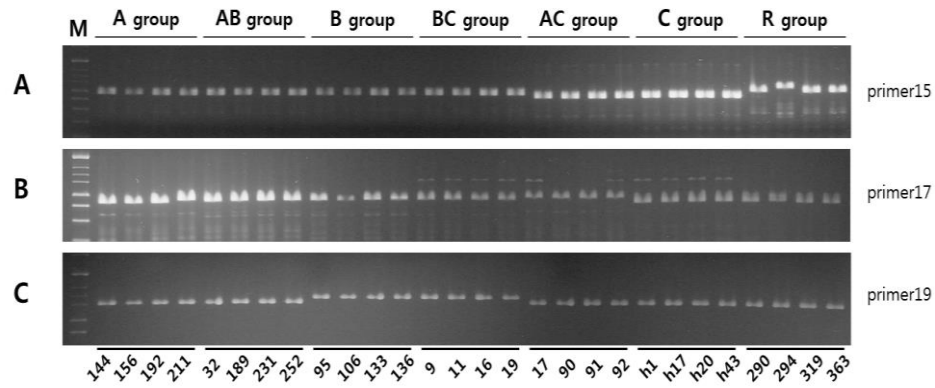

Figure 20. Validation of indel markers for 28 *Brassica* species. (A) PCR amplification using a primer set No. 15, which is AC, C, R group and sample 294 specific. (B) PCR amplification using primer set No. 17, which is specific to AC, C, R and sample 211 in A group. (C) PCR amplification using a primer set No. 19, which is B and BC group specific.

DISCUSSION

The genome sequence of an organism provides the basis for gene discovery, the analysis of genetic variation, and the association of genomic variation with heritable traits. Genome sequence variation can vary from SNPs, insertions/deletions to presence/absence of large regions or rearrangements. Second generation sequencing technologies and applied bioinformatics tools can provide an unprecedented insight into genome structure and variation, with applications for understanding the evolution of *Brassica* species and advancing crop breeding strategies.

Explosive evolution of *Brassica* genus makes *Brassica* species as a potential source for understanding various but important evolutionary aspects such as polyploidy evolution and the impact of genome triplication on morphological evolution of the plants genome (Wang, 2013, Huang et al. 2016). Moreover, the most studied diploid plant *A. thaliana*, belongs to the same family and serve as a bridge to understand the functions of the *Brassica* genome. In addition, application to the *Brassica* genome, further increased our understanding and pursuit to delve more into the genome. Due to the simple and intact form, Chloroplast genomes are a valuable tool for rapid understanding the phylogenetic history of plants species (Moore et al. 2010, Carbonell-Caballero et al. 2015). Compared with other genomes (nuclear DNA, mitochondrial DNA), the CpDNA genomes are highly stable with less mutation rate, and produce highly reliable phylogenetic tree to know the evolutionary history (Mandáková and Lysak, 2008, Nikiforova et al. 2013, Palmer et al. 1983, Yang et al. 2015). Even though the nuclear genomes are prone to cross-hybridization and intermingled during the meiosis, the nrDNA can be maintain at highly homozygous state (Kim et al. 2015a). However, no comprehensive study has been made to understand the diversity and evolution of the genus *Brassica*. In this study, a

comprehensive analysis of the *Brassica* genus to understand the genetic diversity, phylogenetic relationship, and evolution based on complete chloroplast genomes and nrDNA sequences (Kim et al. 2015b) was presented.

Complete CpDNA and nrDNA sequences of major *Brassica* species listed in the classical U's triangle were produced to completely understand the *Brassica* species. Though reference CpDNA and nrDNA genome has been available for few *Brassica* species, the study has newly developed CpDNA source for BBCC genome and nrDNA for BB, BBCC, AACC, and AABB genomes. Moreover, both sub-genome types were identified in all the three allotetraploid *Brassica* species (AB, AC, BC-genomes). According to the results, both CpDNA and nrDNA are highly conserved in terms of gene structure and gene order. There is also very low variation suggesting that CpDNA and nrDNA genomes are evolutionarily conserved. C genome has high conservation within the CpDNA genome but in contact, R genome showed high divergence, suggesting that the continuous evolution of the R genome. The nrDNA genome showed more variations which were observed in the ITS regions.

Copy numbers of CpDNA and nrDNA vary in different plant genomes (Kim et al. 2015b). Though copy numbers estimated based on average read depth coverage of the CpDNA and nrDNA are comparatively similar within the diploids *Brassica* species, drastic sub genome-specific variations for nrDNA copies of the three tetraploids were identified, up to three fold different for BBCC and AACC suggest that these sub-genome dominance mechanisms may have putative role for genome function (Table 14). Though CpDNA and nrDNA are conserved in all the investigated genomes, a considerable number of structural variants were observed to show genus and species specific variations. Those variations are important for understanding the genetic diversity, genome evolution and development

of barcoding markers (Kim et al. 2015b). Barcoding markers based on CpDNA and nrDNA are highly valuable in cultivar identification and authentication. A comprehensive analysis of variants of nrDNA based on seven species leads to development of barcoding marker for discrimination of each species. Only 23 variants (including 22 SNV and an indel) were identified based on 40 different types of nrDNA sequences from 28 accessions (Table 13). Interestingly, analysis of highly potential variants leads to differentiation of A, B, C, and R-genome by single PCR analysis suggesting the importance of nrDNA for barcoding marker development. The 100% homozygosity was observed for the intra-species level, and suggest the high conservation of the nrDNA. However, both inter and intra species diversity based on CpDNA suggest that continuous evolution of the chloroplast genome. Marker validation of CpDNA variants leads to identification of species-specific barcoding markers. Furthermore, among the diploid radish genome shows high intra-diversity suggesting that it undegoes continuous evolution. Furthermore, among the tetraploids, *B. napus* showed members of species-specific variants which are even diverse with their progenitor genomes. Overall, variants based on CpDNA and nrDNA are highly valuable for discrimination of *Brassica* species.

CpDNA analysis showed the relationship between the wild and cultivated citrus genome and the maternal sources of the current cultivars, which are important for further crop improvement (Carbonell-Caballero et al. 2015). The availability of complete CpDNA and nrDNA has certainly enabled us to develop high-quality phylogenetic tree, which eventually provide clear understanding of the genetic relationship of the *Brassica* genus and species of U's triangle. Phylogenetic analysis based on CpDNA and nrDNA shows tmonophyletic origin and clearly demonstrate the genetic relationship of the major diploid and tetraploid *Brassica* species. This holds true especially for nrDNA based phylogeny which has shown exact

association between tetraploids and diploids. However, phylogeny based on CpDNA showed that ambiguous relationship of AC genome with this expected parental genomes (A and C). This enigmatic relationship of *B. napus* chloroplast has been previously investigated (Qiao et al. 2016). The *B. napus* contains three different types of chloroplast genome in its cytoplasm; rap-type, ole-type, and nap-type. Among them rap- and ole-types are close to their progenitors A- and C-genome, respectively, but nap-type is unique and the most abundant (>90% of the B. napus based on 488 accession) present in the B. napus genome. The nap-types are expected to originate from various factors such as source of uninvestigated or wild relatives or natural/artificial selection or introgression (Qiao et al. 2016). The phylogenetic analysis of AC genome based on CpDNA formed independent cluster suggesting the mystery of the parental origin for AC-genome especially for the cytoplasmic source. Furthermore, phylogeny based on nrDNA following their parental trend, supports the A and C as progenitors and strenthen the possibility of different cytoplasmic origin.

Brassica genus is one of the important model plant families for almost all the contemporary plant genomic studies. The exact or reliable temporal timeline of evolution of *Brassica* species is essential for understanding their evolutionary contexts and unifies the hypothesis developed from various disciplines of plant research (Franzke et al. 2016). Various molecular dating approaches (molecular clock and synonymous approach) has been implemented for estimation of divergence times of Brassicaceae (Franzke et al. 2016, Huang et al. 2016, Lysak et al. 2016, Yang et al. 2006a, Arias et al. 2014, Ermolaeva et al. 2003, Lysak et al. 2005, Koch et al. 2001, Hohmann et al. 2015). However, the estimated age is not clear and exhibited different divergence time in accordance with the materials or approach used (Kumar and Hedges, 2016, Huang et al. 2016, Franzke et al. 2016). For example, Brassicaceae divergence yielded about

17-20 MYA and 10-35 MYA based on molecular clock and synonymous approach, respectively (Huang et al. 2012).

Chapter Ⅳ. Conclusion

U’s triangle of *Brassica* includes six major important cultivated *Brassica* species, serve as an important source of oil, vegetables, and fodders worldwide. Despite being extensively studied owing its commercial importance, *Brassica* species has been largely unresolved in terms of diversity, origin, and evolution.

This research elucidates the diversity and evolution of *Brassica* species using 28 re-sequencing data. The genome wide variation in *Brassica* genus were investigated by mapping 28 re-sequenced data to reference genome of *B. rapa*, *B. oleracea,* and *B. napus* and then, the diversity and evolution of the *Brassica* genus were studied among/between those species in U’s triangle. In addition, the complete chloroplast genome sequences of 28 *Brassica* species were generated and by using *de novo* assembly, phylogenetic relationship, and chloroplast structure were analyzed.

This independent divergence analysis based on CpDNA and nrDNA provide a clade understanding of evolutionary timeline for the *Brassica* genus. Tree topology was almost identical with the CpDNA and nrDNA, and clear divergence of each species. Among the five clades based on CpDNA, *B. nigra* was formed as an oldest divergence and estimated around 11.6-5.4 MYA. Independent divergence was observed for R-genome about 11.6-5.4 MYA and conferred before the formation of oleracea lineage was formed around 5.4-2.7 MYA. Furthermore, molecular divergence of allotetraploids reveals that AB, AC, and BC-genomes showed expected divergence rate of 0.3-0.01 MYA. Molecular dating based on nrDNA proved the divergence rate as 0.03-0.001, which is consistent with previous reports. The evolutionary time line of U’s triangle *Brassica* species based on this study and previous reports are summarized in Figure 19.

Structural variants such as SNP and indel have provided potential barcoding markers for identification of each *Brassica* species including

Raphanus sativus. Notably, SNP variants can be used to develop barcoding SNP chip for *Brasscia* species.

This study not only provide insights into *Brassica* genome evolution but also underpin research into the many important crops in this genus and there by helpful to the breeders and geneticists research on *Brassica* species. It is hoped that this modest compendium marks the beginning of a vibrant future for *Brassica* comparative genome biology, gene discovery, molecular development, and genetic dissection of important traits.

REFERENCES

Al-Shehbaz I, Beilstein M, Kellogg E (2006) Systematics and phylogeny of the Brassicaceae (Cruciferae): an overview. *Plant Systematics and Evolution 259* (2-4):89-120

Allen G, Flores-Vergara M, Krasynanski S, Kumar S, Thompson W (2006) A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. *Nature protocols 1* (5):2320-2325

Allender, C. J. & King, G. J. 2010. Origins of the amphiploid species *Brassica napus* L. investigated by chloroplast and nuclear molecular markers. *BMC plant biology, 10,* 54.

Arias, T., Beilstein, M. A., Tang, M., Mckain, M. R. & Pires, J. C. 2014. Diversification times among *Brassica* (Brassicaceae) crops suggest hybrid formation after 20 million years of divergence. *American Journal of Botany, 101,* 86-91.

Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis* thaliana inbred lines. *Nature 465* (7298):627-631

Ayele M, Haas BJ, Kumar N, Wu H, Xiao Y, Van Aken S, Utterback TR, Wortman JR, White OR, Town CD (2005) Whole genome shotgun sequencing of *Brassica oleracea* and its application to gene discovery and annotation in *Arabidopsis*. *Genome Research 15* (4):487-495

Bailey CD, Koch MA, Mayer M, Mummenhoff K, O'Kane SL, Warwick SI, Windham MD, Al-Shehbaz IA (2006) Toward a global phylogeny of the Brassicaceae. *Molecular Biology and Evolution 23* (11):2142-2160

Bauer DC (2011) Variant calling comparison CASAVA1. 8 and GATK.

Bayly, M. J., Rigault, P., Spokevicius, A., Ladiges, P. Y., Ades, P. K., Anderson, C., Bossinger, G., Merchant, A., Udovicic, F. & Woodrow, I. E. (2013). Chloroplast genome analysis ofAustralian eucalypts–Eucalyptus, Corymbia, Angophora, Allosyncarpia and Stockwellia (Myrtaceae). *Molecular Phylogenetics and Evolution 69*, 704–716.

Beecher CW (1994) Cancer preventive properties of varieties of *Brassica oleracea*: a review. *American Journal of Clinical Nutrition 59* (5):1166S-1170S

Birky, C. W. 1995. Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. Proceedings of the National Academy of Sciences, 92, 11331-11338.

Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *The Plant Cell 16* (7):1679-1691

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics:btu170

Bollina V, Khedikar Y, Clarke WE, Parkin IA (2015) 6 Genomics of *Brassica* Oilseeds. *Brassica* Oilseeds: Breeding and Management: 91

Carbonell-Caballero, J., Alonso, R., Ibañez, V., Terol, J., Talon, M. & Dopazo, J. 2015. A phylogenetic analysis of 34 chloroplast genomes elucidates the relationships between wild and domestic species within the genus Citrus. *Molecular Biology and Evolution*, 32.

Cardone M, Mazzoncini M, Menini S, Rocco V, Senatore A, Seggiani M, Vitolo S (2003) *Brassica carinata* as an alternative oil crop for the production of biodiesel in Italy: agronomic evaluation, fuel production by transesterification and characterization. *Biomass and Bioenergy 25* (6):623-636

Chalhoub B, Denoeud F, Liu S, Parkin IA, Tang H, Wang X, Chiquet J, Belcram H, Tong C, Samans B (2014) Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science 345* (6199):950-953

Chen ZJ, Birchler JA (2013) Polyploid and hybrid genomics. Wiley Online Library,

Cheng F, Liu S, Wu J, Fang L, Sun S, Liu B, Li P, Hua W, Wang X (2011) BRAD, the genetics and genomics database for *Brassica* plants. *BMC Plant Biology 11* (1):1

Cheng F, Mandáková T, Wu J, Xie Q, Lysak MA, Wang X (2013) Deciphering the diploid ancestral genome of the mesohexaploid *Brassica rapa*. *The Plant Cell 25* (5):1541-1554

Cheng, F., Wu, J. & Wang, X. 2014. Genome triplication drove the diversification of *Brassica* plants. *Horticulture Research, 1*, 14024.

Chevre A, Eber F, This P, Barret P, Tanguy X, Brun H, Delseny M, Renard M (1996) Characterization of *Brassica nigra* chromosomes and of blackleg resistance in B. napus–B. nigra addition lines. *Plant Breeding 115* (2):113-118

Choi I-Y, Hyten DL, Matukumalli LK, Song Q, Chaky JM, Quigley CV, Chase K, Lark KG, Reiter RS, Yoon M-S (2007) A soybean transcript map: gene distribution, haplotype and single-nucleotide polymorphism analysis. *Genetics 176* (1):685-696

Cronn R, Liston A, Parks M, Gernandt DS, Shen R, Mockler T (2008) Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Research 36* (19):e122-e122

Dassanayake M, Oh D-H, Haas JS, Hernandez A, Hong H, Ali S, Yun D-J, Bressan RA, Zhu J-K, Bohnert HJ (2011) The genome of the extremophile crucifer Thellungiella parvula. *Nature Genetics 43* (9):913-918

Dong, W., Liu, J., Yu, J., Wang, L. & Zhou, S. (2012). Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. PLoS ONE 7, e35071.

Doorduin, L., Gravendeel, B., Lammers, Y., Ariyurek, Y., Chin-A-Woeng, T. & Vrieling, K. (2011). The complete chloroplast genome of 17

individuals of pest species jacobaea vulgaris: SNPs, microsatellites and barcoding markers for population and phylogenetic studies. *DNA Research 18*, 93–105.

Dos Santos J, Nienhuis J, Skroch P, Tivang J, Slocum M (1994) Comparison of RAPD and RFLP genetic markers in determining genetic similarity among *Brassica oleracea* L. genotypes. *Theoretical and Applied Genetics 87* (8):909-915

Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology Evolution 29,* 1969-73.

Ermolaeva, M. D., Wu, M., Eisen, J. A. & Salzberg, S. L. 2003. The age of the *Arabidopsis thaliana* genome duplication. *Plant Molecular Biology, 51,* 859-866.

Feltus FA, Wan J, Schulze SR, Estill JC, Jiang N, Paterson AH (2004) An SNP resource for rice genetics and breeding based on subspecies indica and japonica genome alignments. *Genome Research 14* (9):1812-1819

Franzke, A., Koch, M. A. & Mummenhoff, K. 2016. Turnip Time Travels: Age Estimates in Brassicaceae. Trends in plant science.

Franzke, A., Lysak, M. A., Al-Shehbaz, I. A., Koch, M. A. & Mummenhoff, K. 2011. Cabbage family affairs: the evolutionary history of Brassicaceae. *Trends in Plant Science, 16,* 108-116.

Gaur R, Meena RSR (2016) Multiple disease resistance in different *Brassica* genotypes. *Journal of Oilseed Brassica 1* (1):98-105

Gulick P, Kaczmarek M, Koczyk G, Ziolkowski PA, Babula-Skowronska D, Sadowski J (2009) Comparative analysis of the *Brassica oleracea* genetic map and the *Arabidopsis thaliana* genome. *Genome 52* (7):620-633

Gregory, T. R. (2005). DNA barcoding does not compete with taxonomy. *Nature 434*, 1067.

Hafidh RR, Abdulamir AS, Bakar FA, Jalilian FA, Jahanshiri F, Abas F, Sekawi Z (2013) Novel anticancer activity and anticancer mechanisms of *Brassica oleracea* L. var. capitata f. rubra. *European Journal of Integrative Medicine 5* (5):450-464

Halkier BA, Gershenzon J (2006) Biology and biochemistry of glucosinolates. *Annual Review of Plant Biology 57*:303-333

Hall BG (2013) Building phylogenetic trees from molecular data with MEGA. Molecular biology and evolution:mst012

Hallden C, Nilsson N-O, Rading I, Säll T (1994) Evaluation of RFLP and RAPD markers in a comparison of *Brassica napus* breeding lines. *Theoretical and Applied Genetics 88* (1):123-128

Hasterok, R., Jenkins, G., Langdon, T., Jones, R. N. & Maluszynska, J. 2001. Ribosomal DNA is an effective marker of *Brassica* chromosomes. *Theoretical and Applied Genetics, 103*, 486-490.

Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, Forczek E, Joly-Lopez Z, Steffen JG, Hazzouri KM (2013) An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nature Genetics 45* (8):891-898

Hayward A, Mason A, Dalton-Morgan J, Zander M, Edwards D, Batley J (2012) SNP discovery and applications in *Brassica napus*. *Plant Biotechnology 39* (1):12

Hebert, P. D. N., Penton, E. H., Burns, J. M., Janzen, D. H. & Hallwachs, W. (2004). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. Proceedings of the National Academy of Sciences of the United States of America 101, 14812–14817

Henry RJ (2012) Next-generation sequencing for understanding and accelerating crop domestication. *Briefings in functional genomics 11* (1):51-56

Hohmann, N., Wolf, E. M., Lysak, M. A. & Koch, M. A. 2015. A time-calibrated road map of Brassicaceae species radiation and evolutionary history. *The Plant Cell, 27*, 2770-2784.

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng J-F, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nature Genetics 43* (5):476-481

Hu, Z.-Y., Hua, W., Huang, S.-M. & Wang, H.-Z. 2011. Complete chloroplast genome sequence of rapeseed (*Brassica napus L.*) and its evolutionary implications. *Genetic Resources and Crop Evolution, 58*, 875-887.

Hua W, Li RJ, Zhan GM, Liu J, Li J, Wang XF, Liu GH, Wang HZ (2012) Maternal control of seed oil content in *Brassica napus*: the role of silique wall photosynthesis. *The Plant Journal 69* (3):432-444

Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature Genetics 42* (11):961-967

Huang, C.-C., Hung, K.-H., Wang, W.-K., Ho, C.-W., Huang, C.-L., Hsu, T.-W., Osada, N., Hwang, C.-C. & Chiang, T.-Y. 2012. Evolutionary rates of commonly used nuclear and organelle markers of *Arabidopsis* relatives (Brassicaceae). *Gene, 499*, 194-201.

Huang, C.-H., Sun, R., Hu, Y., Zeng, L., Zhang, N., Cai, L., Zhang, Q., Koch, M. A., Al-Shehbaz, I. & Edger, P. P. 2016. Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. *Molecular Biology and Evolution, 33*, 394-412.

Huang, C. Y., Gr¨unheit, N., Ahmadinejad, N., Timmis, J. N. & Martin, W. (2005). Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. *Plant Physiology 138*, 1723–1733.

Izzah NK, Lee J, Jayakodi M, Perumal S, Jin M, Park B-S, Ahn K, Yang T-J (2014) Transcriptome sequencing of two parental lines of cabbage (*Brassica oleracea* L. var. capitata L.) and construction of an EST-based genetic map. *BMC Genomics 15* (1):1

Johnston JS, Pepper AE, Hall AE, Chen ZJ, Hodnett G, Drabek J, Lopez R, Price HJ (2005) Evolution of genome size in Brassicaceae. *Annals of Botany 95* (1):229-235

Kagale S, Koh C, Nixon J, Bollina V, Clarke WE, Tuteja R, Spillane C, Robinson SJ, Links MG, Clarke C (2014) The emerging biofuel crop Camelina sativa retains a highly undifferentiated hexaploid genome structure. Nature communications 5

Kim, K., Lee, S.-C., Lee, J., Lee, H. O., Joh, H. J., Kim, N.-H., Park, H.-S. & Yang, T.-J. 2015a. Comprehensive Survey of Genetic Diversity in Chloroplast Genomes and 45S nrDNAs within *Panax ginseng* Species. *PLoS ONE*, 10, e0117159.

Kim, K., Lee, S.-C., Lee, J., Yu, Y., Yang, K., Choi, B.-S., Koh, H.-J., Waminal, N. E., Choi, H.-I. & Kim, N.-H. 2015b. Complete chloroplast and ribosomal sequences for 30 accessions elucidate evolution of Oryza AA genome species. Scientific reports, 5.

Koch M, Kiefer C (2006) Molecules and migration: biogeographical studies in cruciferous plants. *Plant Systematics and Evolution 259* (2-4):121-142

Koch, M., Haubold, B. & Mitchell-Olds, T. 2001. Molecular systematics of the Brassicaceae: evidence from coding plastidic matK and nuclear Chs sequences. *American Journal of Botany, 88*, 534-544.

Koenig D, Weigel D (2015) Beyond the thale: comparative genomics and genetics of *Arabidopsis* relatives. *Nature Reviews Genetics 16* (5):285-298

Koo, D. H., Hong, C. P., Batley, J., Chung, Y. S., Edwards, D., Bang, J. W., Hur, Y. & Lim, Y. P. 2011. Rapid divergence of repetitive DNAs in *Brassica* relatives. *Genomics, 97*, 173-85.

Kopsell DA, Kopsell DE (2006) Accumulation and bioavailability of dietary carotenoids in vegetable crops. *Trends in Plant Science 11* (10):499-507

Kumar, S. & Hedges, S. B. 2016. Advances in Time Estimation Methods for Molecular Data. *Molecular Biology and Evolution, 33*, 863-869.

Kumar, S., Stecher, G. & Tamura, K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology Evolution*.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods 9* (4):357-359

Lee J, Izzah NK, Jayakodi M, Perumal S, Joh HJ, Lee HJ, Lee S-C, Park JY, Yang K-W, Nou I-S (2015) Genome-wide SNP identification and

QTL mapping for black rot resistance in cabbage. *BMC Plant Biology 15* (1):1

Lee T-H, Guo H, Wang X, Kim C, Paterson AH (2014) SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics 15* (1):162

Li G, Quiros CF (2001) Sequence-related amplified polymorphism (SRAP), a new marker system based on a simple PCR reaction: its application to mapping and gene tagging in *Brassica*. *Theoretical and Applied Genetics 103* (2-3):455-461

Li M, Zhang C, Liu L, Yu L (2005) Development of relationship between A, B and C genomes in *Brassica* genera. *Hereditas* 27 (4):671-676

Li, X., Yang, Y., Henry, R. J., Rossetto, M., Wang, Y. & Chen, S. 2015. Plant DNA barcoding: from gene to genome. *Biological Reviews, 90*, 157-166.

Lim YP, Plaha P, Choi SR, Uhm T, Hong CP, Bang JW, Hur YK (2006) Toward unraveling the structure of *Brassica rapa* genome. *Physiologia Plantarum 126* (4):585-591

Lim, K. B., De Jong, H., Yang, T. J., Park, J. Y., Kwon, S. J., Kim, J. S., Lim, M. H., Kim, J. A., Jin, M., Jin, Y. M., Kim, S. H., Lim, Y. P., Bang, J. W., Kim, H. I. & Park, B. S. 2005. Characterization of rDNAs and tandem repeats in the heterochromatin of *Brassica rapa*. *Molecules and Cells, 19*, 436-44.

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z (2012) GAPIT: genome association and prediction integrated tool. *Bioinformatics 28* (18):2397-2399

Liu, S., Liu, Y., Yang, X., Tong, C., Edwards, D., Parkin, I. A., Zhao, M., Ma, J., Yu, J., Huang, S., Wang, X., Wang, J., Lu, K., Fang, Z., Bancroft, I., Yang, T. J., Hu, Q., Wang, X., Yue, Z., Li, H., Yang, L., Wu, J., Zhou, Q., Wang, W., King, G. J., Pires, J. C., Lu, C., Wu, Z., Sampath, P., Wang, Z., Guo, H., Pan, S., Yang, L., Min, J., Zhang, D., Jin, D., Li, W., Belcram, H., Tu, J., Guan, M., Qi, C., Du, D., Li, J., Jiang, L., Batley, J., Sharpe, A. G., Park, B. S., Ruperao, P., Cheng, F., Waminal, N. E., Huang, Y., Dong, C., Wang, L., Li, J., Hu, Z., Zhuang, M., Huang, Y., Huang, J., Shi, J., Mei, D., Liu, J., Lee, T. H., Wang, J., Jin, H., Li, Z., Li, X., Zhang, J., Xiao, L., Zhou, Y., Liu, Z., Liu, X., Qin, R., Tang, X., Liu, W., Wang, Y., Zhang, Y., Lee, J., Kim, H. H., Denoeud, F., Xu, X., Liang, X., Hua, W., Wang, X., Wang, J., Chalhoub, B. & Paterson, A. H. 2014. The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. *Nature Communications, 5*, 3930.

Lohse, M., Drechsel, O. & Bock, R. 2007. OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Current Genetics, 52*, 267-274.

Lukens LN, QUIJADA PA, UDALL J, Pires JC, Schranz M, Osborn TC (2004) Genome redundancy and plasticity within ancient and recent *Brassica* crop species. *Biological Journal of the Linnean Society 82* (4):665-674

Luo, H., Shi, J., Arndt, W., Tang, J. & Friedman, R. (2008). Gene order phylogeny of the genus Prochlorococcus. *PLoS ONE* 3, e3837.

Luo, H., Sun, Z., Arndt, W., Shi, J., Friedman, R. & Tang, J. (2009). Gene order phylogeny and the evolution of methanogens. *PLoS ONE* 4, e6069.

Lysak, M. A., Koch, M. A., Pecinka, A. & Schubert, I. 2005. Chromosome triplication found across the tribe Brassiceae. *Genome Research, 15*, 516-525.

Lysak, M. A., Mandáková, T. & Schranz, M. E. 2016. Comparative paleogenomics of crucifers: ancestral genomic blocks revisited. *Current Opinion in Plant Biology, 30*, 108-115.

Mandáková, T. & Lysak, M. A. 2008. Chromosomal phylogeny and karyotype evolution in x= 7 crucifer species (Brassicaceae). *The Plant Cell, 20*, 2559-2570.

McNally KL, Childs KL, Bohnert R, Davidson RM, Zhao K, Ulat VJ, Zeller G, Clark RM, Hoen DR, Bureau TE (2009) Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proceedings of the National Academy of Sciences* 106 (30):12273-12278

McPherson, H., van der Merwe, M., Delaney, S. K., Edwards, M. A., Henry, R. J., McIntosh, E., Rymer, P. D., Milner, M. L., Siow, J. & Rossetto, M. (2013). Capturing chloroplast variation for molecular ecology

studies: a simple next generation sequencing approach applied to a rainforest tree. *BMC Ecology 13*, 8.

Michael TP, Jackson S (2013) The first 50 plant genomes. *The Plant Genome 6* (2)

Moore MJ, Dhingra A, Soltis PS, Shaw R, Farmerie WG, Folta KM, Soltis DE (2006) Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biology 6* (1):1

Moore, M. J., Soltis, P. S., Bell, C. D., Burleigh, J. G. & Soltis, D. E. 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proceedings of the National Academy of Sciences*, 107, 4623-4628.

Mun J-H, Kwon S-J, Yang T-J, Seol Y-J, Jin M, Kim J-A, Lim M-H, Kim JS, Baek S, Choi B-S (2009) Genome-wide comparative analysis of the *Brassica rapa* gene space reveals genome shrinkage and differential loss of duplicated genes after whole genome triplication. *Genome Biology 10* (10): 111

U N(1935) Genome analysis in *Brassica* with special reference to the experimental formation of B. napus and peculiar mode of fertilization. *The Journal of Japanese Botany 7*:389-452

Negi M, Sabharwal V, Bhat S, Lakshmikumaran M (2004) Utility of AFLP markers for the assessment of genetic diversity within *Brassica nigra* germplasm. *Plant Breeding 123* (1):13-16

Nikiforova, S. V., Cavalieri, D., Velasco, R. & Goremykin, V. 2013. Phylogenetic analysis of 47 chloroplast genomes clarifies the contribution of wild species to the domesticated apple maternal line. *Molecular Biology and Evolution, 30,* 1751-1760.

Ohyama K, Fukuzawa H, Kohchi T, Shirai H, Sano T, Sano S, Umesono K, Shiki Y, Takeuchi M, Chang Z (1986) Chloroplast gene organization deduced from complete sequence of liverwort Marchantia polymorpha chloroplast DNA. *Nature 322*:572-574

Pakpour S, Klironomos J (2015) The invasive plant, *Brassica nigra*, degrades local mycorrhizas across a wide geographical landscape. *Royal Society Open Science 2* (9):150300

Parks, M., Cronn, R. & Liston, A. (2009). Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biology 7*, 84–100.

Palmer, J. D., Shields, C., Cohen, D. & Orton, T. 1983. Chloroplast DNA evolution and the origin of amphidiploid *Brassica* species. *Theoretical and Applied Genetics, 65,* 181-189.

Park J, Koo D, Hong C, Lee S, Jeon J, Lee S, Yun P, Park B, Kim H, Bang J (2005) Physical mapping and microsynteny of *Brassica rapa* ssp. pekinensis genome corresponding to a 222 Kbp gene-rich region of *Arabidopsis* chromosome 4 and partially duplicated on chromosome 5. *Molecular Genetics and Genomics 274* (6):579-588

Park S, Yu H-J, Mun J-H, Lee S-C (2010) Genome-wide discovery of DNA polymorphism in *Brassica rapa*. *Molecular Genetics and Genomics 283* (2):135-145

Parkin IA, Koh C, Tang H, Robinson SJ, Kagale S, Clarke WE, Town CD, Nixon J, Krishnakumar V, Bidwell SL (2014) Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica oleracea*. *Genome Biology 15* (6):1-18

Parkin, I. A., Koh, C., Tang, H., Robinson, S. J., Kagale, S., Clarke, W. E., Town, C. D., Nixon, J., Krishnakumar, V., Bidwell, S. L., Denoeud, F., Belcram, H., Links, M. G., Just, J., Clarke, C., Bender, T., Huebert, T., Mason, A. S., Pires, J. C., Barker, G., Moore, J., Walley, P. G., Manoli, S., Batley, J., Edwards, D., Nelson, M. N., Wang, X., Paterson, A. H., King, G., Bancroft, I., Chalhoub, B. & Sharpe, A. G. 2014. Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica oleracea*. *Genome Biology*, *15*, R77.

Paterson AH, Lan T-h, Amasino R, Osborn TC, Quiros C (2001) *Brassica* genomics: a complement to, and early beneficiary of, the *Arabidopsis* sequence. *Genome Biology 2* (3):1339-1347

Piquemal J, Cinquin E, Couton F, Rondeau C, Seignoret E, Perret D, Villeger M-J, Vincourt P, Blanchard P (2005) Construction of an oilseed rape (*Brassica napus L.*) genetic map with SSR markers. *Theoretical and Applied Genetics 111* (8):1514-1523

Pradhan A, Gupta V, Mukhopadhyay A, Arumugam N, Sodhi Y, Pental D (2003) A high-density linkage map in *Brassica juncea* (Indian mustard) using AFLP and RFLP markers. *Theoretical and Applied Genetics 106* (4):607-614

Qian W, Meng J, Li M, Frauen M, Sass O, Noack J, Jung C (2006) Introgression of genomic components from Chinese *Brassica rapa* contributes to widening the genetic diversity in rapeseed (B. napus L.), with emphasis on the evolution of Chinese rapeseed. *Theoretical and Applied Genetics 113* (1):49-54

Qiao, J., Cai, M., Yan, G., Wang, N., Li, F., Chen, B., Gao, G., Xu, K., Li, J. & Wu, X. 2016. High-throughput multiplex CpDNA resequencing clarifies the genetic diversity and genetic relationships among *Brassica napus, Brassica rapa* and *Brassica oleracea*. *Plant Biotechnology Journal, 14*, 409-418.

Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics. *Current Opinion in Plant Biology 5* (2):94-100

Rahman M, McVetty PB, Li G (2007) Development of SRAP, SNP and Multiplexed SCAR molecular markers for the major seed coat color gene in *Brassica rapa* L. *Theoretical and Applied Genetics 115* (8):1101-1107

Reboud, X. & Zeyl, C. 1994. Organelle inheritance in plants. *Heredity, 72*, 132-140.

Schattner, P., Brooks, A. N. & Lowe, T. M. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Research, 33*, W686-W689.

Seol, Y.-J., Kim, K., Kang, S.-H., Perumal, S., Lee, J. & Kim, C.-K. 2015. The complete chloroplast genome of two *Brassica* species, *Brassica nigra* and *B. oleracea*. *Mitochondrial DNA*, 1-2.

Seol, Y.-J., Lee, T.-H., Park, D.-S.& Kim, C.-K. 2016. NABIC: A New Access Portal to Search, Visualize, anhare Agricultural Genomics Data. *Evolutionary Bioinformatics*, 2016:12 51–58

Sharma A, Li X, Lim YP (2014) Comparative genomics of Brassicaceae crops. *Breeding Science 64* (1):3

Sharma, S., Padmaja, K. L., Gupta, V., Paritosh, K., Pradhan, A. K. & Pental, D. 2014. Two plastid DNA lineages—Rapa/Oleracea and Nigra—within the tribe Brassiceae can be best explained by reciprocal crosses at hexaploidy: evidence from divergence times of the plastid genomes and R-block genes of the A and B genomes of *Brassica juncea*. *PloS ONE*, 9, e93260.

Shekhawat K, Rathore S, Premi O, Kandpal B, Chauhan J (2012) Advances in agronomic management of Indian mustard (*Brassica juncea* (L.) Czernj. Cosson): an overview. *International Journal of Agronomy* 2012

Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchi-

Shinozaki K (1986). The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *The EMBO Journal 5* (9):2043

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics 15* (2):121-132

Singh R, SINGH AK, KUMAR P (2014) Performance of Indian Mustard (*Brassica juncia* L.) in Response to Integrated Nutrient Management. *Journal of AgriSearch 1* (1)

Slotte T, Hazzouri KM, Å gren JA, Koenig D, Maumus F, Guo Y-L, Steige K, Platts AE, Escobar JS, Newman LK (2013) The Capsella rubella genome and the genomic consequences of rapid mating system evolution. *Nature Genetics 45* (7):831-835

Struss D, Quiros C, Plieske J, Röbbelen G (1996) Construction of *Brassica* B genome synteny groups based on chromosomes extracted from three different sources by phenotypic, isozyme and molecular markers. *Theoretical and Applied Genetics 93* (7):1026-1032

Subbaiyan GK, Waters DL, Katiyar SK, Sadananda AR, Vaddadi S, Henry RJ (2012) Genome-wide DNA polymorphisms in elite indica rice inbreds discovered by whole-genome sequencing. *Plant Biotechnology Journal 10* (6):623-634

Szewc-McFadden A, Kresovich S, Bliek S, Mitchell S, McFerson J (1996) Identification of polymorphic, conserved simple sequence repeats

(SSRs) in cultivated *Brassica* species. *Theoretical and Applied Genetics 93* (4):534-538

Tanhuanpää P, Vilkki J, Vilkki H (1995) Association of a RAPD marker with linolenic acid concentration in the seed oil of rapeseed (*Brassica napus* L.). *Genome 38* (2):414-416

Town CD, Cheung F, Maiti R, Crabtree J, Haas BJ, Wortman JR, Hine EE, Althoff R, Arbogast TS, Tallon LJ (2006) Comparative genomics of *Brassica oleracea* and *Arabidopsis thaliana* reveal gene loss, fragmentation, and dispersal after polyploidy. *The Plant Cell 18* (6):1348-1359

Trick M, Long Y, Meng J, Bancroft I (2009) Single nucleotide polymorphism (SNP) discovery in the polyploid *Brassica napus* using Solexa transcriptome sequencing. *Plant Biotechnology Journal 7* (4):334-346

Varshney, R. K. & May, G. D. 2012. Next-generation sequencing technologies: opportunities and obligations in plant genomics. *Brief Functional Genomics*, *11*, 1-2.

Venglat P, Xiang D, Yang H, Wan L, Tibiche C, Ross A, Wang E, Selvaraj G, Datla R (2013) Gene expression profiles during embryo development in *Brassica napus*. *Plant Breeding 132* (5):514-522

Waminal, N. E., Perumal, S., Lim, K.-B., Park, B.-S., Kim, H. H. & Yang, T.-J. 2015. Genomic Survey of the Hidden Components of the *B. rapa* Genome. *The Brassica rapa Genome*. Springer.

Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun J-H, Bancroft I, Cheng F (2011a) The genome of the mesopolyploid crop species *Brassica rapa*. *Nature Genetics 43* (10):1035-1039

Wang Y, Sun S, Liu B, Wang H, Deng J, Liao Y, Wang Q, Cheng F, Wang X, Wu J (2011b) A sequence-based genetic linkage map as a reference for *Brassica rapa* pseudochromosome assembly. *BMC Genomics 12* (1):1

Wang Y, Wang X, Paterson AH (2012) Genome and gene duplications and gene expression divergence: a view from plants. *Annals of the New York Academy of Sciences 1256* (1):1-14

Wang, X. 2013. The *Brassica* genome, Frontiers E-books.

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., Bai, Y., Mun, J. H., Bancroft, I., Cheng, F., Huang, S., Li, X., Hua, W., Freeling, M., Pires, J. C., Paterson, A. H., Chalhoub, B., Wang, B., Hayward, A., Sharpe, A. G., Park, B. S., Weisshaar, B., Liu, B., Li, B., Tong, C., Song, C., Duran, C., Peng, C., Geng, C., Koh, C., Lin, C., Edwards, D., Mu, D., Shen, D., Soumpourou, E., Li, F., Fraser, F., Conant, G., Lassalle, G., King, G. J., Bonnema, G., Tang, H., Belcram, H., Zhou, H., Hirakawa, H., Abe, H., Guo, H., Jin, H., Parkin, I. A., Batley, J., Kim, J. S., Just, J., Li, J., Xu, J., Deng, J., Kim, J. A., Yu, J., Meng, J., Min, J., Poulain, J., Hatakeyama, K., Wu, K., Wang, L., Fang, L., Trick, M., Links, M. G., Zhao, M., Jin, M., Ramchiary, N., Drou, N., Berkman, P. J., Cai, Q., Huang, Q., Li, R., Tabata, S., Cheng, S., Zhang, S., Sato, S., Sun, S., Kwon, S. J., Choi, S. R., Lee, T. H., Fan, W., Zhao, X., Tan, X.,

Xu, X., Wang, Y., Qiu, Y., Yin, Y., Li, Y., Du, Y., Liao, Y., Lim, Y., Narusaka, Y., Wang, Z., Li, Z., Xiong, Z. & Zhang, Z. 2011. The genome of the mesopolyploid crop species *Brassica rapa*. *Nature Genetics, 43*, 1035-9.

Warwick, S. I., Mummenhoff, K., Sauder, C. A., Koch, M. A. & Al-Shehbaz, I. A. 2010. Closing the gaps: phylogenetic relationships in the Brassicaceae based on DNA sequence data of nuclear ribosomal ITS region. *Plant Systematics and Evolution, 285*, 209-232.

Wyman, S. K., Jansen, R. K. & Boore, J. L. 2004. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics, 20*, 3252-3255.

Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nature Biotechnology 30* (1):105-111

Yang R, Jarvis DE, Chen H, Beilstein MA, Grimwood J, Jenkins J, Shu S, Prochnik S, Xin M, Ma C (2013) The reference genome of the halophytic plant Eutrema salsugineum. *Front Plant Science 4* (46): 10

Yang T-J, Kim JS, Kwon S-J, Lim K-B, Choi B-S, Kim J-A, Jin M, Park JY, Lim M-H, Kim H-I (2006) Sequence-level analysis of the diploidization process in the triplicated FLOWERING LOCUS C region of *Brassica rapa*. *The Plant Cell 18* (6):1339-1347

Yang, J. B., Tang, M., Li, H. T., Zhang, Z. R. & Li, D. Z. (2013). Complete chloroplast genome of the genus Cymbidium: lights into the species

identification, phylogenetic implications and population genetic analyses. *BMC Evolutionary Biology 13*, 84.

Yang, J., Liu, G., Zhao, N., Chen, S., Liu, D., Ma, W., Hu, Z. & Zhang, M. 2015. Comparative mitochondrial genome analysis reveals the evolutionary rearrangement mechanism in *Brassica*. *Plant Biology*, n/a-n/a.

Yu J, Zhao M, Wang X, Tong C, Huang S, Tehrim S, Liu Y, Hua W, Liu S (2013) Bolbase: a comprehensive genomics database for *Brassica oleracea*. *BMC Genomics 14* (1):1

Zhang, Y., Du, L., Liu, A., Chen, J., Wu, L., Hu, W., Zhang, W., Kim, K., Lee, S.-C. & Yang, T.-J. 2016. The complete chloroplast genome sequences of five Epimedium species: lights into phylogenetic and taxonomic analyses. *Frontiers in Plant Science, 7*.

국문 초록 (Abstract in Korean)

배추속(*Brassica* genus) 식물은 모델 식물인 애기장대(*A. thaliana*)와 2천만년 전에 분화하였고, 농업적으로 중요한 다양한 작물을 포함한다. 농업적으로 중요한 배추속 6종들은 이배체(diploid)인 배추 (*B. rapa*, AA, 2n = 20), 양배추 (*B. oleracea*, CC, 2n = 18) 그리고 흑겨자 (*B. nigra*, BB, 2n = 16)가 있고, 이들 사이의 자연적인 종간교잡에 의해서 탄생한 이질사배체(Allotetraploid)인, 유채 (*B. napus*, AACC, 2n = 38), 갓 (*B. juncea*, AABB, 2n = 34) 그리고 에티오피아겨자(*B. carinata*, BBCC, 2n = 36)가 있다.

배추과 종들의 진화 및 다양성을 이해하기 위해 배추과 유전체 A, B, C, R, AB, AC 타입에 해당하는 28개체에 대해 일루미나 사의 MiSeq 차세대 시퀀싱 플랫폼을 이용하여 유전체 재 분석을 수행하였다. 총 각 개체별로 6~8백만개의 서열을 생산하였다. 유전체 해독이 끝나 공개된 배추, 양배추, 유체 및 애기장대를 레퍼런스로 맵핑을 수행하여 개체별로 레퍼런스 지놈크기의 평균 3배(3x depth coverage) 정도의 짧은 서열들이 맵핑되었고 총 약 5천9백만개의 단일염기다형성(SNP)과 약 2만4천개의 삽입-결실(indel)을 보였다. 배추과 종보다는 애기장대의 유전자간(Intergenic) 지역 부분의 SNP 가 상대적으로 매우 적게 분포했으며, 이는 진화적으로 유전자간 지역의 변이가 훨씬 많아 유전자간 지역에 맵핑(mapping)이 안되어 나타나는 현상으로 보인다. 맵핑된 SNP 정보를 기반으로 지놈연관분석을 수행하였으며, 이배체 개체들은 레퍼런스 지놈에 상관없이 종끼리 군집을 이루었다. 반면 이질사배체의 개체들은 다소 부정확하고 레퍼런스 지놈에 따라 다른 양상을 보였다. 이는 배추과 식물의 배수체화(polyploid)등의 복잡한 진화 진화에

기인하는 것으로 보인다. 반면 진화적으로 거리가 먼 애기장대를 레퍼런스로 이질사배체를 포함하여 정교한 SNP로 분석한 결과 배추과 식물의 진화와 종의 합성에 대한 연관분석이 좀더 분명하게 분석 되었다.

또한, 배추과 28개체에 대해 엽록체 지놈(CpDNA)과 45S 라이보좀 서열(nrDNA)을 조립하고 이를 통해 계통발생학적(phylogenetic) 분석을 수행하여 배추과 식물의 다양성과 진화에 관한 분석을 하였다. 계통발생학적(phylogenetic) 분석을 통해 배추과 식물의 유전적 다양성과 종간 연관성 및 진화 그리고 이질사배체의 모계 종을 밝힐 수 있다. 또한, 엽록체 지놈과 45S 라이보좀 서열에 대한 단일염기다형성 및 삽입-결실 변이에 대한 전체 지도를 구축하였다. 엽록체 지놈과 45S 라이보좀 서열에 대한 변의를 바탕으로 이배체의 분화시기와 이질사배체의 탄생 시기를 측정하였다. 엽력체 지놈과 45S 라이보좀 서열에 대한 변의를 활용해 무를 포함한 배추과 식물의 종 동정을 위한 바코드용 마커를 개발하고 이를 검증하였다. 엽록체 지놈과 라이보좀 서열 분석은 배추과 식물의 다양성과 진화 연구에 종합적인 해석을 가능하게 하였다.

주요어: 전장 유전체 재분석, 차세대염기서열분석, 단일염기서열다형성, 엽록체 염기서열, 라이보좀 염기서열, 배추과

학 번: 2009-30107